computation

MDPI

Article

Intelligent Real-Time Deep System for Robust Objects Tracking in Low-Light Driving Scenario

Francesco Rundo

STMicroelectronics, ADG Central R&D, Zona Industriale, 95121 Catania, Italy; francesco.rundo@st.com; Tel.: +39-335-7367811

Abstract: The detection of moving objects, animals, or pedestrians, as well as static objects such as road signs, is one of the fundamental tasks for assisted or self-driving vehicles. This accomplishment becomes even more difficult in low light conditions such as driving at night or inside road tunnels. Since the objects found in the driving scene represent a significant collision risk, the aim of this scientific contribution is to propose an innovative pipeline that allows real time low-light driving salient objects tracking. Using a combination of the time-transient non-linear cellular networks and deep architectures with self-attention, the proposed solution will be able to perform a real-time enhancement of the low-light driving scenario frames. The downstream deep network will learn from the frames thus improved in terms of brightness in order to identify and segment salient objects by bounding-box based approach. The proposed algorithm is ongoing to be ported over a hybrid architecture consisting of a an embedded system with SPC5x Chorus MCU integrated with an automotive-grade system based on STA1295 MCU core. The performances (accuracy of about 90% and correlation coefficient of about 0.49) obtained in the experimental validation phase confirmed the effectiveness of the proposed method.

Keywords: assisted driving; intelligent driving scenario understanding; deep learning; low-light self-driving; low-light driving saliency detection

Citation: Rundo, F. Intelligent Real-Time Deep System for Robust Objects Tracking in Low-Light Driving Scenario. *Computation* **2021**, *9*, 117. https://doi.org/10.3390/computation9110117

Academic Editors: Yudong Zhang and Demos T. Tsahalis

Received: 19 October 2021
Accepted: 3 November 2021
Published: 8 November 2021

1. Introduction

Autonomous Driving (AD) or ADAS, i.e., Advanced Driver Assisting Systems, are considered very promising technology/based solutions able to cover the safety requirements in such very complex automotive scenarios [1,2]. Both for an autonomous vehicle and assisted driving, it is critical to have a normal-light captured driving video-frames as most of the classical computer vision algorithms degrade significantly in the absence of adequate lighting [3].

The lack of sufficient illumination of the driving scene does not allow the semantic segmentation algorithms to identify and track objects or the deep classifiers to discriminate one class rather than another due to poor significant pixel-based information and therefore limited discriminating visual features [3]. Autonomous driving (AD) and driver assistance (ADAS) systems require high levels of robustness both in performance and fault-tolerance, often requiring high levels of validation and testing before being placed on the market [4]. The author has already deeply investigated the main issues and critical points of the ADAS technologies [4–10]. Specifically, in the contributions reported in [4–7] such drowsiness detection methods based on the analysis of the car driver physiological signals have been reported. An innovative method named "hyper-filtering" combined with robust computer vision algorithms has been reported in [8–10] to provide an intelligent assistance to the car driver.

Therefore, the light-level of the driving scenario could have a significant impact on the object detection task, even though the recent proposed approaches are not able to address the main target of autonomous or assisted driving in poor light environment. For these

reasons, several scientific studies investigated the "perception" in self-driving technology in a driving scenario with a poor illumination. A very robust perception system includes a LiDAR based approach with an embedded visual camera [4–10]. As introduced, the performance of the car's camera device is significantly affected by the level of light as well as low contrast, low visibility, and so on. As will be described in the next section, the proposed pipelines have made some progress in processing low-light driving images even though there are still several drawbacks such as poor effects on background, light interference, rain, and snow induced noise, etc. Moreover, the time consumption of traditional approaches is often too long to get a real-time processing. With the aim of addressing the mentioned issues, the author has deeply investigated the problem. Specifically, the design of a robust and effective approach was investigated with aim to improve the discriminating features of low-illumination visual driving frames with a near real-time processing.

This scientific contribution is organized as follows: the next section "Related works" includes the analysis of the state-of-the-art in the field of intelligent solutions for low-light driving scenario enhancement while Section 3 "Materials and Methods" describes in detail the proposed solution. Finally, the Section 4 includes the outcome of the experimental assessment of the proposed approach while Section 5 "Discussion and Conclusion" includes such description of the main advantages of the proposed pipeline with some ideas for the future development.

2. Related Work

As introduced in the previous section, the aim of the methods reported in scientific literature suitable to address the problem of driving at low light conditions embeds the common task to provide an enhancement of driving video frames in times compatible with automotive constraints.

For this purpose, the traditional image processing approaches used for early image enhancement were preliminarily used, but the related performances were often limited. After the emergence of machine and deep learning methods, the researchers exploited these new algorithmic technologies which included Deep Convolutional Neural Networks (CNNs), Recurrent Neural Networks, Generative Model Networks, and similar in order to perform complex and adaptive video/image processing tasks.

More in detail, the flowchart followed by the scientific community was to apply the most recent deep learning-based techniques of image and video processing to characterize the frames of low-light driving scenarios. Deep ad-hoc networks have been implemented, to leverage the capability of the already implementing pre-trained deep architectures. As introduced in the work, in some cases, some researchers have integrated to the visual images other data coming from different sensors such as RADAR or LiDAR. Here are some significant scientific works in relation to the aforementioned techniques.

In [11] an interesting approach has been proposed. The authors proposed an approach to enhance such type of low-light images through regularized illumination optimization combined with a pipeline for noise suppression. Based on Retinex theory, the authors implemented two deep architectures (E-Net and D-Net) for noise level estimation and noise reduction, respectively. The performance confirmed the effectiveness of the proposed approach [11].

With specific focus on the artificial light enhancement approaches applied to video and image processing, an interesting method was proposed by Qu et al., in [12]. They proposed a deep network based on the usage of a Cycle Generative Adversarial Networks (CycleGAN) combined with additional discriminators. They tested their solution for addressing the task of the autonomous robot navigation retrieving very promising results.

In [13] the authors analyzed an interesting framework based on bio-inspired solutions. The authors analyzed the bio-inspired solutions based on the working-flow of the human retina. They proposed an event-based neuromorphic vision system suitable to convert asynchronous driving events into synchronous 2D representation more efficient for object detection and tracking applications. Ad-hoc light sensitive cells which contain millions

of photoreceptors combined with deep learning have been proposed in [13] to overcome the low-light driving issues. The experimental results they retrieved were promising and deserve further study.

The authors in [14] designed a deep network named FuseMODNet to cover the issue of low-light driving scenario in AD and ADAS framework. They proposed a robust and very promising Deep architecture for Moving Object Detection (MOD) under poor light condition. They obtained in testing session a promising 10.1% relative improvement on common Dark-KITTI dataset, and a 4.25% improvement on standard KITTI dataset [14].

An interesting approach for addressing the analysis of low-light video frames is based on the so called Retinex decomposition. [3,15]. More details in the next paragraphs.

In [15] the authors used a so called "Retinex" decomposition-based solution for low-light image enhancement with joint decomposition and denoising. They have applied a denoising algorithm through the usage of ad-hoc trained customized U-Net network applied to a properly composed low-light images dataset. Experimental results based on LOL dataset confirmed the effectiveness of the pipeline proposed.

In [3] Pham et al. investigated the usage of Retinex theory as effective tool for enhancing the illumination and detail of images. They collected a Low-Light Drive (LOL-Drive) dataset and applied a deep retinex neural network, named Drive-Retinex, which was taught using this dataset. The deep Retinex-Net consists of two subnetworks: Decom-Net (decomposes a color image into a reflectance map and an illumination map) and Enhance-Net (enhances the light level in the illumination map). Their experimental sessions confirmed that the proposed method was able to achieve visually appealing low-light enhancement.

Such authors have investigated the issue of low-light image analysis in automotive applications. A first solution was proposed by Szankin et al. in [16]. They analyzed the application of low-power system for road surface classification and pedestrian detection in challenging environments, including low-light driving. The authors investigated the impact of such external features (lightning conditions, moisture of the road surface and ambient temperature) on the system ability to properly detect the pedestrian and road in the driving scenario. The implemented system was tested on images captured in different climate zones. The solution they reported in [16] based on the usage of deep network showed very promising results as the pedestrian detection tests confirmed precision and recall above 95% in challenging driving scenario. More details in [16].

In [17] another interesting approach based on retinex theory was presented. The authors of [17] designed an end-to-end intelligent architecture combined with a data-driven mapping network with layer-specified constraints for single-image low-light enhancement. A Sparse Gradient Minimization sub-Network (SGM-Net) was implemented to remove the low-amplitude structures and preserve major edge information. Two sub deep networks (Enhance-Net and Restore-Net) were configured to predict the enhanced illumination and reflectance maps, respectively, which helps stretch the contrast of the illumination map and remove intensive noise in the reflectance map. The so structured pipeline showed very promising effectiveness which significantly outperforms the state-of-the-art methods.

Although the analyzed methods allow us to obtain excellent results in enhancing video frames associated with a low-brightness driving scenario, they often fail to find an optimal trade-off between accuracy, performance, and speed of execution [3,14–17]. The method herein proposed tries to balance the above items, providing robust and acceptable performance in a near real-time and sustainable hardware framework. The next section will introduce and detail the proposed pipeline.

3. Materials and Methods

As introduced, the aim of this proposal is the design of an innovative system that allows addressing the issue of assisted or autonomous driving in low light driving scenario conditions. This algorithm must be sustainable for the underlying hardware as well as guarantee near real-time response times. Preliminarily, the implemented pipeline will

be introduced and therefore the individual sub-systems will be detailed in the related sub-sections The proposed pipeline is schematized in Figure 1.

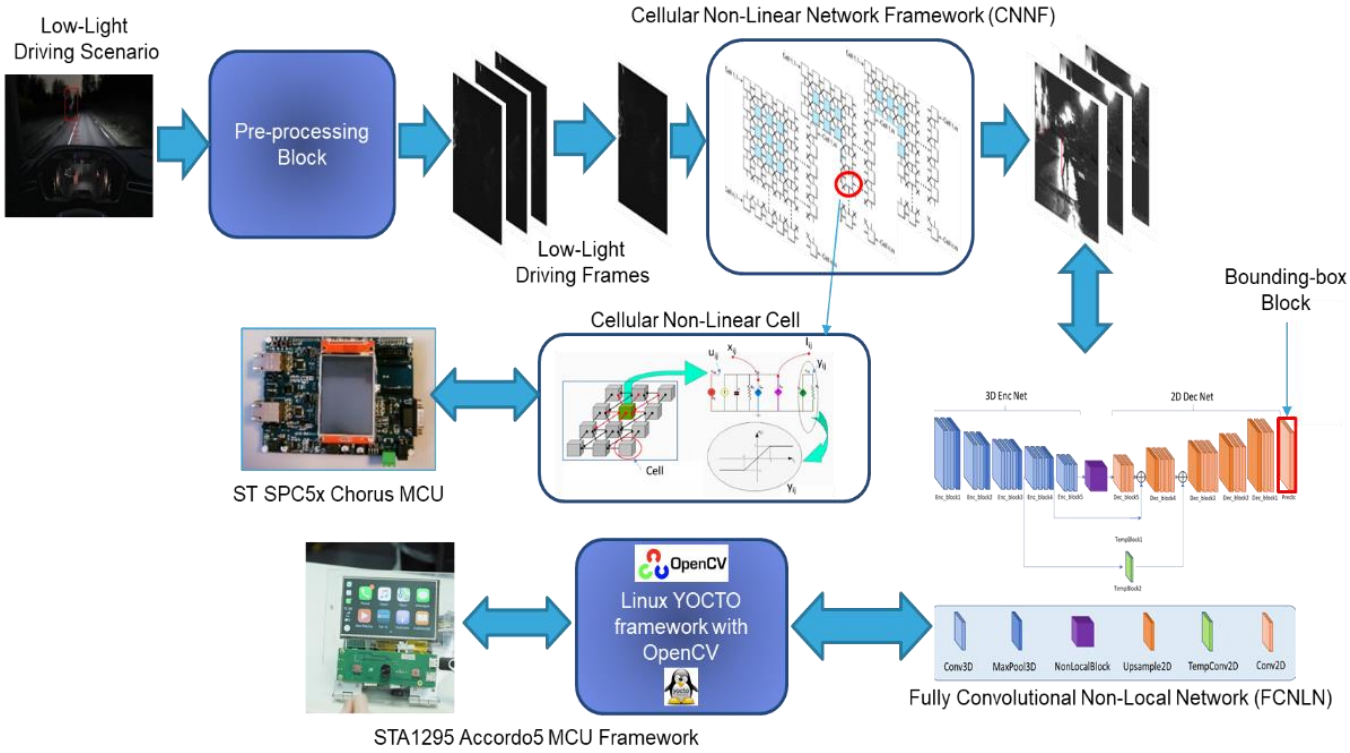


Figure 1. The schematic overview of the proposed pipeline.

As schematized in Figure 1, the proposed pipeline is composed by three sub-systems: the first sub-system is the "Pre-processing block", the second one is the "Cellular Non-Linear Network Framework", and finally there is the "Fully Convolutional Non-Local Network". Each of the implemented sub-systems will be described in the next sections.

3.1. The Pre-Processing Block

As schematized in Figure 1, the first sub-system of the proposed pipeline is the pre-processing block. The target of this block is the normalization of the sampled visual frames that represent the driving scenario. Specifically, the sampled low-light driving frames coming from the used low-frame-rate camera (60 fps) will be resized (through classic bi-cubic algorithm) to a size of 256 × 256. Any type of low frame-rate (40–60 fps) camera with a minimum resolution of 640 × 480 can be used with reference to the proposed method. A resizing of 256 × 256 was chosen as a trade-off between information content of the sampled driving frames and computational load of the embedded system in which the algorithm will be carried.

The external automotive-grade camera will be a gray-level device and, in case of color camera, the gray level image will be extracted from a classic YCbCr conversion. The single frame $D_k(x,y)$ thus pre-processed will be buffered in the memory area of the micro SPC5x MCU [8–11] and gradually processed by the Cellular Non-Linear Network Framework as described in the next paragraph.

3.2. The Cellular Non-Linear Network Framework

As schematized in Figure 1, the normalized pre-processed low-light driving frames $D_k(x,y)$ will be further processed by the designed Cellular Non-Linear Network Framework. The goal of this block is the artificial increasing of the brightness level of the frames representative of the driving scenario in order to get a robust object detection or segmenta-

tion. Specifically, in the designed framework a revision of the classical Cellular Non-Linear architecture has been proposed, i.e., an architecture which shows a time-transient dynamic. A brief introduction about the Time-transient Cellular Non-linear Network (TCNN) is then reported.

The first architecture of the Cellular Neural (or Nonlinear) Network (CNN) was firstly designed by L.O Chua and L. Yang [18,19]. The CNN backbone is basically a high-speed local arranged computing array of analog processing units called "cells" [19]. Many applications and extensions of the original CNN architecture have been proposed in literature [20,21]. The basic unit of the CNNs is the cell. The CNNs dynamic can be defined through the instructions embedded in the related cloning template configuration matrices [18,19]. Each cell of the CNN array evolves as dynamical temporal system spatially arranged into a topological 2D or 3D representation. The CNN cells interact each other within its neighborhood configured by heuristically mathematical model. Each CNN cell has an input, a state, and an output variable which is a functional mapping of the state (usually a simple PieceWise Linear dynamic).

The CNNs architecture can be implemented with analog circuits or by means of U-VLSI technology performing near real-time analogic processing task. Some dynamic-stability results of the CNNs can be found in [22]. An updated version of original Chua-Yang CNN model was introduced by Arena et al. in [22–24] and called "State Controlled Cellular Neural Network (SC-CNN)" as it directly correlates dynamic state-evolution of the cell.

However, in the classic CNN paradigm such specific space-time dependency between state, input and neighborhood can be modelized. Consequently, the CNN can be designed as a dynamical system of cells (or neurons) defined on a normed space S_N (cell grid), discrete subset of $\Re^n$ (generally $n \leq 3$) with ad-hoc metric function d: $S_N - > N$ (N: integer set). Cells are identified by indices defined in a space-set L_i. Neighborhood function $Nr(k)$ of the k-th cell, can be defined as follows:

$$N_r : L_i \rightarrow L_i^{\beta} \tag{1}$$

$$N_r(k) = \{l | d(i,j) \leq r\} \tag{2}$$

where β depends on r (neighborhood radius) and on space geometry of the grid. Cells are multiple input–single output nonlinear processors. The cell dynamic is defined by its state, which is generally not observable outside of the cell itself. As an introduction, the author remarks that every CNNs cell is only connected to other cells within the defined neighborhood.

The CNNs can have a structure with as single layer or multi-layers having a spatial arrangement such as a planar array (with rectangular, square, octagonal geometry) or a k-dimensional array (usually $k \geq 3$), generally considered and realized as a stack of k-dimensional arrays (layers). In the defined spatial structure, the CNN can embeds identical cells or mixed solutions (as is the case for neurons), with different neighborhood size, variables, etc.

To summarize what was described: a CNN can be defined as time-continuous—space-discrete system represented by the contemporary spatio-temporal dynamics of the single cells whose differential state-model maybe represented as follows:

$$\begin{aligned} \frac{\partial x_j}{\partial t} = g[x_j(t)] + & \sum_{\gamma \in N_r(j)} \aleph_{\vartheta_j}\left(x_j\Big|_{(t-\tau,t]}, y_\gamma\Big|_{(t-\tau,t]}; p_j^A\right) \\ & + \sum_{\gamma \in N_r(j)} \mathcal{B}_{\varphi_j}\left(x_j\Big|_{(t-\tau,t]}, u_\gamma\Big|_{(t-\tau,t]}; p_j^B\right) \\ & + \sum_{\gamma \in N_r(j)} \mathbb{C}_{\rho_j}\left(x_j\Big|_{(t-\tau,t]}, x_\gamma\Big|_{(t-\tau,t]}; p_j^C\right) \\ & + I_j(t) \\ y_j(t) = \Im & \left(x_j\Big|_{(t-\tau,t]}\right) \end{aligned} \tag{3}$$

In Equation (3), x_j, y_j, u_j, I_j denote respectively cell state, output, input, and bias; j and γ are cell indices; $g[x_j(t)]$ is a local instantaneous feedback function; Nr is neighborhood function; p_j^A, p_j^B, p_j^C are arrays of custom configurable parameters; notation x|T denotes the restriction of function x(•) to interval T of its argument; $\aleph_{\vartheta_j}$ is neighborhood feedback functional, and in the same way $\mathcal{B}_{\varphi_j}$ is input functional; $\mathbb{C}_{\rho_j}$ is cell-state functional; and $\Im$ is output functional.

For the pipeline herein described, a further extension of the CNN was used by using a new cloning template matrix $D(i,j;k,l)$ as follows:

$$\begin{aligned} C\frac{\mathrm{dx}_{ij}(t,t_k)}{\mathrm{dt}} = & -\tfrac{1}{R_x}x_{ij} \\ & + \sum_{C(k,l)\in N_r(i,j)} A(i,j;k,l)y_{kl}(t,t_k) \\ & + \sum_{C(k,l)\in N_r(i,j)} B(i,j;k,l)u_{kl}(t,t_k) \\ & + \sum_{C(k,l)\in N_r(i,j)} C(i,j;k,l)x_{kl}(t,t_k) \\ & + \sum_{C(k,l)\in N_r(i,j)} D(i,j;k,l)\left(y_{ij}(t),y_{kl}(t),t_k\right) + I \\ & (1\le i\le M, 1\le j\le N) \\ y_{ij}(t) = & \tfrac{1}{2}\left(\left|x_{ij}(t,t_k)+1\right| - \left|x_{ij}(t,t_k)-1\right|\right) \\ N_r(i,j) = & \{C_r(k,l); (\max(|k-i|,|i-j|)\le r, 1\le k\le M, 1\le l\le N)\} \end{aligned} \tag{4}$$

In Equation (4) the $Nr(i,j)$ represents the neighborhood of each cell $C(i,j)$ with radius r. The terms x_{ij}, y_{kl}, u_{kl}, and I are respectively the state, the output, and the input of the cell $C(i,j)$, while $A(i,j;k,l)$, $B(i,j;k,l)$, $C(i,j;k,l)$, and $D(i,j;k,l)$ are the cloning templates which, as described, are able to configure the space-time evolution of the structured CNNs.

As described by the researchers [19–24] through the setup of the cloning templates (matrix coefficients) as well as the bias I, it is possible to configure the type of processing provided by the CNNs, i.e., the type of features map according to the specific input data.

In this contribution, the author propose a time-transient CNN (TCNN), i.e., a non-linear network which dynamically evolves in a short time range, i.e., during the transient $[t,t_k]$. Normally, CNN evolves up to a defined steady-state [19–22]. The dynamic mathematical model of the TCNN is reported in Equation (4).

Specifically, the input video frames of the TCNN will be the visual 256 × 256 low-light frames in which therefore each cell will be associated with corresponding pixel (intensity) of the selected input visual frame $D_k(x,y)$. Basically, both the CNNs cell state x_{ij} and input u_{ij} will be bound with the corresponding low-light frame pixel intensity of the 256 × 256 image. Each setup of the cloning templates and bias $A(i,j;k,l)$, $B(i,j;k,l)$, $C(i,j;k,l)$, $D(i,j;k,l)$, I will allow us to retrieve a specific processing of the input frame in order to provide an artificial enhancement of the low-light frame.

As introduced in [18–24], there is no analytic method to retrieve the coefficients of the TCNN cloning templates associated to a specific visual task. There are several databases [21–24] but each processing task need ad-hoc cloning templates setup. For this reason, several heuristic-optimization tests have been performed in order to detect the right cloning templates setup suitable to reach the desired target, i.e., the light enhancement of such low-light driving frames.

After several testing sessions, the following cloning templates $A(i,j;k,l)$, $B(i,j;k,l)$, $C(i,j;k,l)$, $D(i,j;k,l)$, and I setup sequentially applied to each of the input pre-processed video frame $D_k(x,y)$ have been identified as very promising for the purpose of the proposed approach. Basically, by means of each of the defined cloning templates setup reported in Equations (5)–(7), the implemented system will be able to improve the light condition of the original frames progressively. The proposed TCNN configuration is described by the following set of mathematical setup:

$$\begin{array}{l} \mathrm{A} = [0.01\ 0.02\ 0.02\ 0.02\ 0.01\ 0.02\ 0.02\ 0.03\ 0.02\ 0.02\ 0.02\ 0.03\ 0.04\ 0.03\ 0.02\ 0.02\ 0.02\ 0.03\ 0.02\ 0.02\ 0.01\ 0.02\ 0.02\ 0.02\ 0.01]; \\ \mathrm{B} = [0.01\ 0.02\ 0.02\ 0.02\ 0.01\ 0.02\ 0.02\ 0.03\ 0.02\ 0.02\ 0.02\ 0.03\ 0.04\ 0.03\ 0.02\ 0.02\ 0.02\ 0.03\ 0.02\ \ 0.02\ 0.01\ 0.02\ 0.02\ 0.02\ 0.01]; \\ \mathrm{C} = 0; \\ \mathrm{D} = 0; \\ \mathrm{I} = -1; \\ \text{Transient Steps (tk)} = 5 \end{array} \tag{5}$$

$$\begin{array}{l} \mathrm{A} = [1\ 0\ 0\ 1\ 4\ -1\ 1\ 0\ 0]; \\ \mathrm{B} = 0; \\ \mathrm{C} = 0; \\ \mathrm{D} = 0; \\ \mathrm{I} = -1; \\ \text{Transient Steps (tk)} = 7 \end{array} \tag{6}$$

$$\begin{array}{l} \mathrm{A} = [1\ 1\ 1\ 1\ 6\ 0\ 1\ 0\ -1]; \\ \mathrm{B} = [0.01\ 0.02\ 0.02\ 0.02\ 0.01\ 0.02\ 0.02\ 0.03\ 0.02\ 0.02\ 0.02\ 0.03\ 0.04\ 0.03\ 0.02\ 0.02\ 0.02\ 0.03\ 0.02\ 0.02\ 0.01\ 0.02\ 0.02\ 0.02\ 0.01]; \\ \mathrm{C} = [1\ 1\ 1\ 0\ 4\ 0\ 0\ -1\ 0]; \\ \mathrm{D} = [0.03\ 0.03\ 0.03\ 0.03\ 0.01\ 0.01\ 0.01\ 0.03\ 0.03\ 0.03\ 0.03\ 0.01\ 0.01\ 0.01\ 0.01\ 0.01\ 0.01\ 0.03\ 0.03\ 0.03\ 0.03\ 0.03\ 0.03\ 0.03\ 0.03]; \\ \mathrm{I} = -0.5; \\ \text{Transient Steps (tk)} = 5 \end{array} \tag{7}$$

In the following Figure 2 we report such instances of the TCNN enhanced input low-light driving frames, with a detail of each light-enhancement generated by each of the TCNN configurations as per Equations (5)–(7).

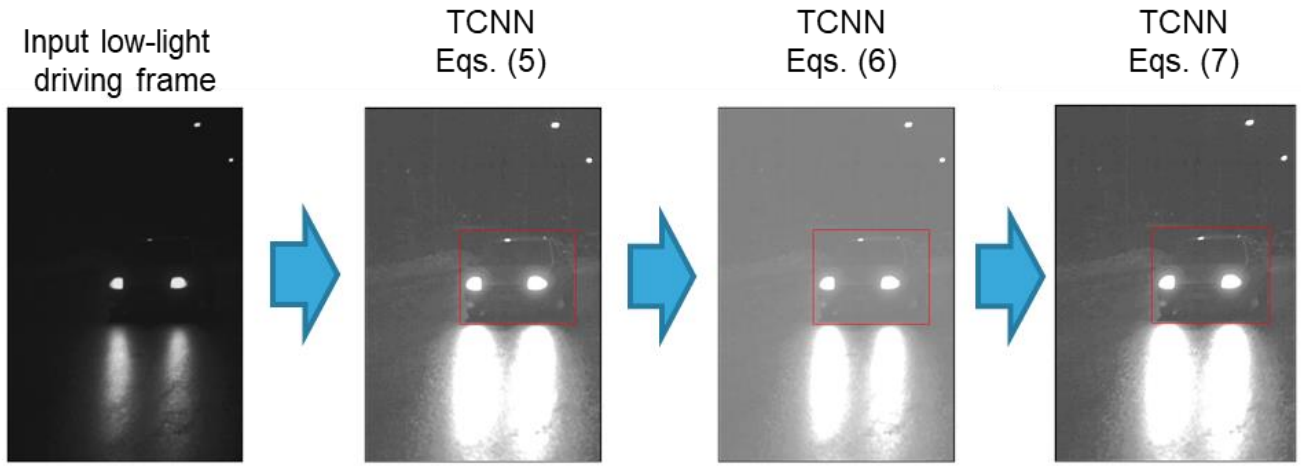


Figure 2. An instance of the low-light image enhancement performed by the proposed TCNN.

As shown in Figure 2, the TCNN-based framework is able to significantly improve the light exposure of the source image representative of the driving scenario. The output image which is represented by the image processed by the cloning templates setup reported in Equation (7), represents the frame on which the downstream intelligent segmentation deep architecture will be applied.

3.3. The Fully Convolutional Non-Local Network

The target of this block is the salient detection and segmentation (though bounding box approach) of the enhanced input driving video frames. As showed in Figure 1, the output of the previous TCNN block will be fed as input to this sub-system, i.e., the Fully Convolutional Non-Local Network (FCNLN). The collected driving scene video frames

will be trained by ad-hoc designed 3D to 2D Semantic Segmentation Fully Convolutional Non-Local Network as reported in Figure 1.

Through a semantic encoding/decoding segmentation of the driving context, the saliency map of the driving scene will be retrieved. This saliency map will be included in the bounding-box block which reconstruct the segmentation Region of Interest (ROI). The proposed FCNLN architecture is structured as follows.

The encoder block (3D Enc Net) extracts the space-time features of the sampled driving frames and it is made up of 5 blocks. The first two blocks include (for each block) two separable convolution layers with $3 \times 3 \times 2$ kernel filter followed by a batch normalization, ReLU layer, and a downstream $1 \times 2 \times 2$ max-pooling layer. The remaining three blocks (for each block) include two separable convolution layers with $3 \times 3 \times 3$ kernel filter followed by a batch normalization, another convolutional layer with $3 \times 3 \times 3$ kernel, batch normalization, and ReLU with a downstream $1 \times 2 \times 2$ max-pooling layer. Before feeding the so processed features to the Decoder side, a Non-Local self attention block is included in the proposed backbone.

Non-local blocks have been recently introduced [25] as a very promising method for leveraging space-time long-range dependencies and correlation to provide more robust features maps [26]. Non-local blocks as self-attention mechanism take inspiration from the non-local means method, extensively applied in computer vision [25,26]. Self-attention through non-local blocks aims to embed a strong and robust correlation among feature maps by ad-hoc properly weighting of those features [25]. In our pipeline, non-local blocks operate between the 3D encoder and 2D decoder side respectively, with the goal to extract features in space-time long-range dependencies embedded in the sampled enhanced driving frames.

The mathematical model of non-local operation is introduced. Given a generic deep architecture and a general input data x, the employed non-local operation determines the corresponding response y_i (of the given network) at i location in the input data as a weighted sum of the input data at all positions $j \neq i$:

$$y_i = \frac{1}{\psi(x)} \sum_{\forall j} \zeta(x_i,\ x_j)\beta(x_j) \tag{8}$$

where $\zeta\ (\cdot)$ is a pairwise potential describing the affinity or relationship between data positions at index i and j respectively. The function $\beta(\cdot)$ is, instead, a unary potential modulating ζ according to input data. The sum is then normalized by a factor $\psi(x)$. The parameters of ζ, β, and ψ potentials are learned during model's training and defined as follows:

$$\zeta(x_i,\ x_j) = e^{\Theta(x_i)^T\ \Phi(x_j)} \tag{9}$$

where Θ and Φ are two linear transformations of the input data x with learnable weights W_Θ and W_Φ:

$$\begin{array}{l} \Theta(x_i) = W_\Theta x_i \\ \Phi(x_j) = W_\Phi x_j \\ \beta(x_j) = \ W_\beta x_j \end{array} \tag{10}$$

For the $\beta(\cdot)$ function, a common linear embedding (classical $1 \times 1 \times 1$ convolution) with learnable weights W_β is employed. The normalization function ψ is:

$$\psi(x) = \sum_{\forall j} \zeta(x_i,\ x_j) \tag{11}$$

The above mathematical formulation of Non-Local features processing is named Embedded Gaussian | [20]. The output of Non-Local processing of the encoded features will be fed in the decoder side of the pipeline.

The Decoder backbone (2D Dec Net) is structured according to the encoder backbone for up-sampling/decode the visual self-attention features of the input encoder. The decoder backbone is composed by five blocks including 2D convolutional layers with 3×3 kernel, batch normalization layers, ReLU. Such residual connections through convolutional block are added. Ad-hoc up-sampling blocks (with bi-cubic algorithm) has been embedded to adjust the size of the feature maps. The output of the so designed FCNLN is the feature saliency map of the sample driving video frame, i.e., the segmented area of the most salient object. The following Figure 3 shows some instance of the saliency output of the proposed FCNLN applied to such video reporting the driving scene.

Figure 3. Saliency analysis of the video representing the driving scene.

The bounding-box block will define the bounding area around the saliency map by means of an enhanced minimum rectangular box criteria (increased by 20% along each dimension) which is able to enclose the salience area. The following Figure 4 shows an example of an automatically generated bounding-box (red rectangle):

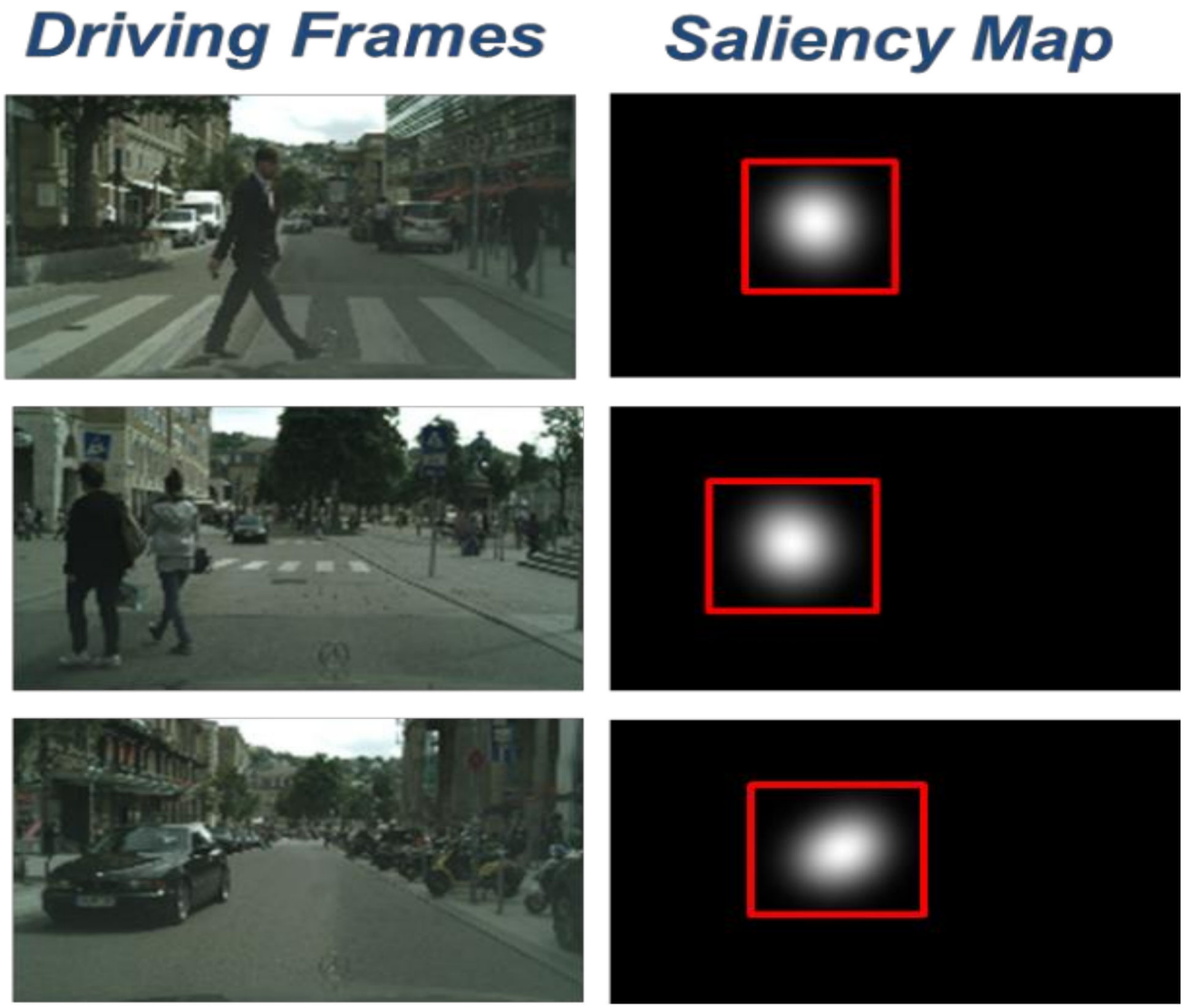


Figure 4. Bounding-box segmentation (red rectangular box) of the saliency map.

The proposed FCNLN architecture has been validated on the DHF1K public dataset [27] retrieving the following performance: Area Under the Curve: 0.899; Similarity: 0.455; Correlation Coefficient: 0.491; Normalized Scanpath Saliency: 2.772.

More robust deep intelligent architectures can be proposed but they would fail in the target to have an embedded automotive-grade portable solution over the proposed hardware (STA1295A Accordo5 and SPC58x Chorus MCU) [8,28,29].

4. Experimental Results

To test the proposed overall system, the same dataset used for similar pipelines has been used. Specifically, for the Video Saliency Scene Understanding Block, the author has extracted images from the following dataset: Oxford Robot Car dataset [30] and the Exclusively Dark (ExDark) Image Dataset [31].

This so composed dataset contains more than 20 million images having an average resolution greater than 640 × 480. We select 15,000 driving frames of that dataset to compose the training set. It contains several complex night low-light driving scenes. Moreover, further testing and validation sessions have been made over several further driving scenario frames we collected from real driving. The author has split the dataset as follow: 70% for training as well as 30% for testing and validation of the proposed approach.

In order to perform a robust validation of the proposed approach, a cross-validation of the designed pipeline has been performed through k-fold (k = 5) approach. The collected experimental results will be reported in Tables 1 and 2.

Table 1. Benchmark comparison with similar pipeline without low-light enhancement.

Pipeline	Number of Detected Objects	Accuracy	NSS	CC
Ground Truth	12,115	/	/	/
Yolo V3	7878	0.650	1.552	0.331
FCN DenseNet-201 backbone	8321	0.686	1.654	0.340
Faster-R-CNN ResNet-50 backbone	8050	0.664	1.614	0.338
Mask-R-CNN ResNet-50 backbone	9765	0.806	2.003	0.377
RetinaNet ResNet50 backbone	9991	0.824	2.45	0.441
Faster-CNN MobileNetv3 backbone	7988	0.650	1.542	0.329
Proposed	**10,937**	**0.902**	**2.654**	**0.488**
Proposed w/o Non-Local Blocks	9.601	0.792	1.998	0.371

Table 2. Benchmark comparison with low-light enhancement.

Pipeline	Number of Detected Objects	Accuracy	NSS	CC
Ground Truth	12,115	/	/	/
Yolo V3	9980	0.823	2.425	0.441
FCN DenseNet201 backbone	9106	0.751	1.857	0.359
Faster-R-CNN ResNet-50 backbone	9230	0.761	1.899	0.362
Mask-R-CNN ResNet-50 backbone	10,115	0.834	2.502	0.461
RetinaNet ResNet50 backbone	10.227	0.844	2.539	0.469
Faster CNN MobileNet v3 backbone	9003	0.743	1.809	0.351
Proposed	**10,937**	**0.902**	**2.654**	**0.488**
Proposed w/o NLocal Blocks & TCNN	8.587	0.708	1.691	0.350

The cloning templates setup of the TCNN is the one reported in Equations (5)–(7) while the FCNLN has been trained with a mini-batch gradient descent with Adam optimizer and initial learning rate of 0.01. The deep model is implemented using Pytorch framework. Experiments were carried out on a server with Intel Xeon CPUs equipped with a Nvidia GTX 2080 GPU with 16 Gbyte ad memory video.

The collected experimental benchmark has reported in the following Table 1 and in which the performance of the whole pipeline is compared with similar deep architecture in which the input is the low-light image. To validate the performance of the proposed pipeline in relation to detecting objects in driving scenarios with poor lighting, benchmark tests were performed against robust architectures known in the scientific community as intelligent object detectors. Specifically, benchmark tests sessions were performed including solutions based on YoloV3, DenseNet-201, ResNet-50, Mask-R-CNN, and RetinaNet (backbone ResNet-50) and Faster-CNN based on MobileNET-v3 as backbone [32]. As performance indicators, accuracy and the following two indices used by the scientific community to validate computer vision algorithms were used [32]:

$$\begin{gathered} NSS(S, G^B) = \frac{1}{N} \sum_i \overline{S_i} \times G_i^B \\ N = \sum_i G_i^B \; ; \overline{S} = \frac{S - \mu(S)}{\sigma(S)} \end{gathered} \tag{12}$$

where G_i^B represents the fixations points on the generated normalized saliency map S i.e., points in which the Ground Truth map G^B is equals to 1 (NSS stands for Normalized Scanpath Saliency)

$$CC(S, G) = \frac{cov(S, G)}{\sigma_S \cdot \sigma_G} \tag{13}$$

where *cov*(*S*,*G*) represents the covariance of normalized saliency map *S* and normalized Ground Truth Map *G* and related standard deviations (σ_S, σ_G). The indicator *CC(S,G)* stands for Correlation Coefficient.

As showed in Table 1, the proposed pipeline outperforms the compared deep architecture in terms of accuracy in object detections. In order to validate the scientific contribution made by TCNN low-light enhancement block, we compared the same architectures as per Table 1 but with input frames coming from the implemented TCNN with a cloning templates setup reported in Equations (5)–(7).

From a visual analysis of the data reported in Table 2, a significant increase in the performance of the comparison pipelines is highlighted, due to the low-light enhancement block based on TCNN.

A comparison of the results reported in Tables 1 and 2 confirmed the excellent performance of the proposed method. Specifically, the analysis of the ablation experiments shows how the TCNN and self-attention Non-Local blocks significantly improve the performance of the system both in object detection and in segmentation skills.

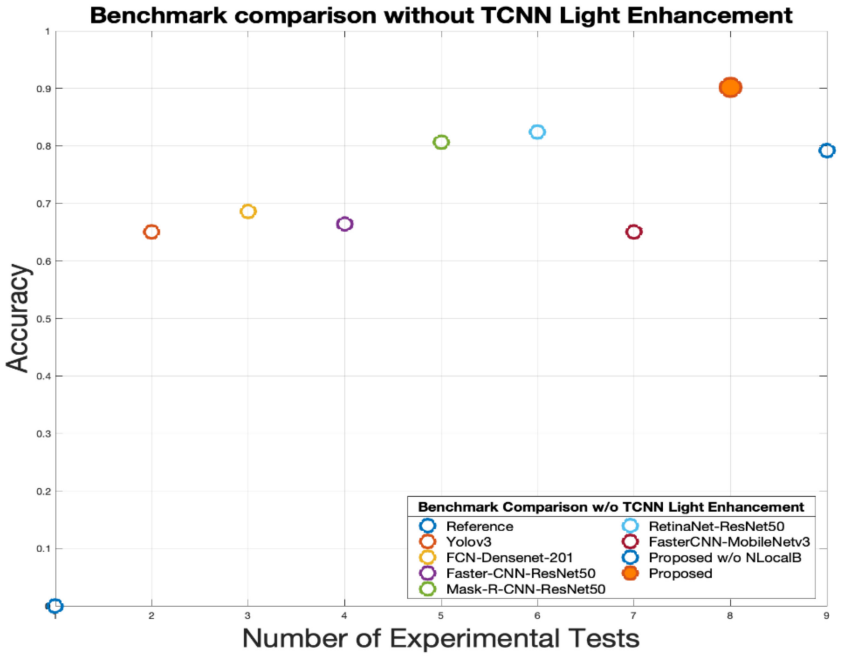
The following diagrams show the benchmark comparison of the compared architectures in terms of accuracy both with TCNN based light enhancement and without.

From Figure 5a,b the need to keep the contribution of the TCNN and self-attention Non-Local blocks is highlighted as their absence significantly degrades the performance of the architecture as it does not allow to determine discriminating features useful for providing the deep downstream network with information for perform object detection and tracking operations. In the automotive field, such drop in performance is not acceptable in terms of driving safety.

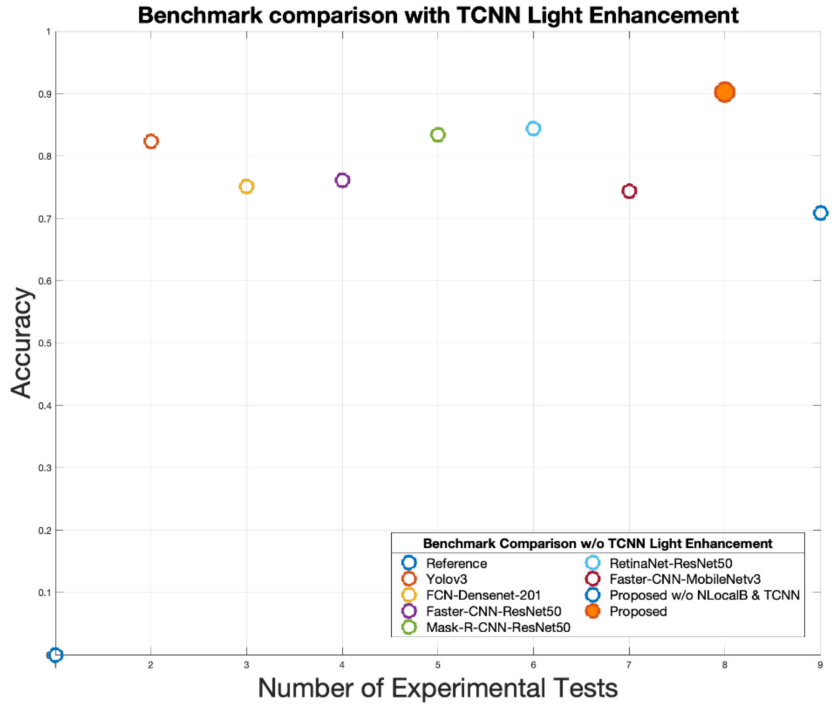
Solutions and backbones have been chosen that are portable in automotive-grade architectures currently in use in the automotive market as well as the MCUs based on SPC58X and STA1295 Accordo5 and that are equipped with the relative ASIL-X certifications. More complex architectures could probably obtain higher performances but are not easily integrable on automotive platforms and therefore go beyond this work which focuses on embedded solutions integrable on automotive frameworks [7–11].

As for timing, all the comparing platforms allow us to obtain a near real-time response compatible with automotive requirements. Specifically, in the preliminary tests we are performing in such prototype version of the proposed pipeline (running in the embedded aforementioned MCUs systems) it is possible to obtain the detection, segmentation and therefore the tracking (in the videos of the driving scenarios) of the objects in a timing less than one/two second/s with a slight superiority of the Faster R-CNN solution with a MobileNet v3, although with less accuracy (see Tables 1 and 2). The proposed pipeline is being ported and optimized in the framework based on SPC5x Chorus MCU which will host the TCNN processing. The thus obtained light-enhanced frames will then be processed by the remaining part of the pipeline shown in Figure 1 and which will be hosted in the STA1295 Accordo5 platform which integrates a 3D graphics accelerator.

The proposed whole pipeline was able to perform better than the others also in terms of classification and this would seem to be attributable to the action of Non-Local self-attention blocks as the network without these blocks degrades in performance (Accuracy 90.672% against 82.558%). The following Figures 5–7 report some instances of the enhanced driving video frames with embedded bounding box of the detected salient objects.



(**a**)



(**b**)

Figure 5. (**a**) Benchmark comparison in terms of accuracy without TCNN low-light enhancement pre-processing; (**b**) benchmark comparison in terms of accuracy with the proposed TCNN low-light enhancement pre-processing.

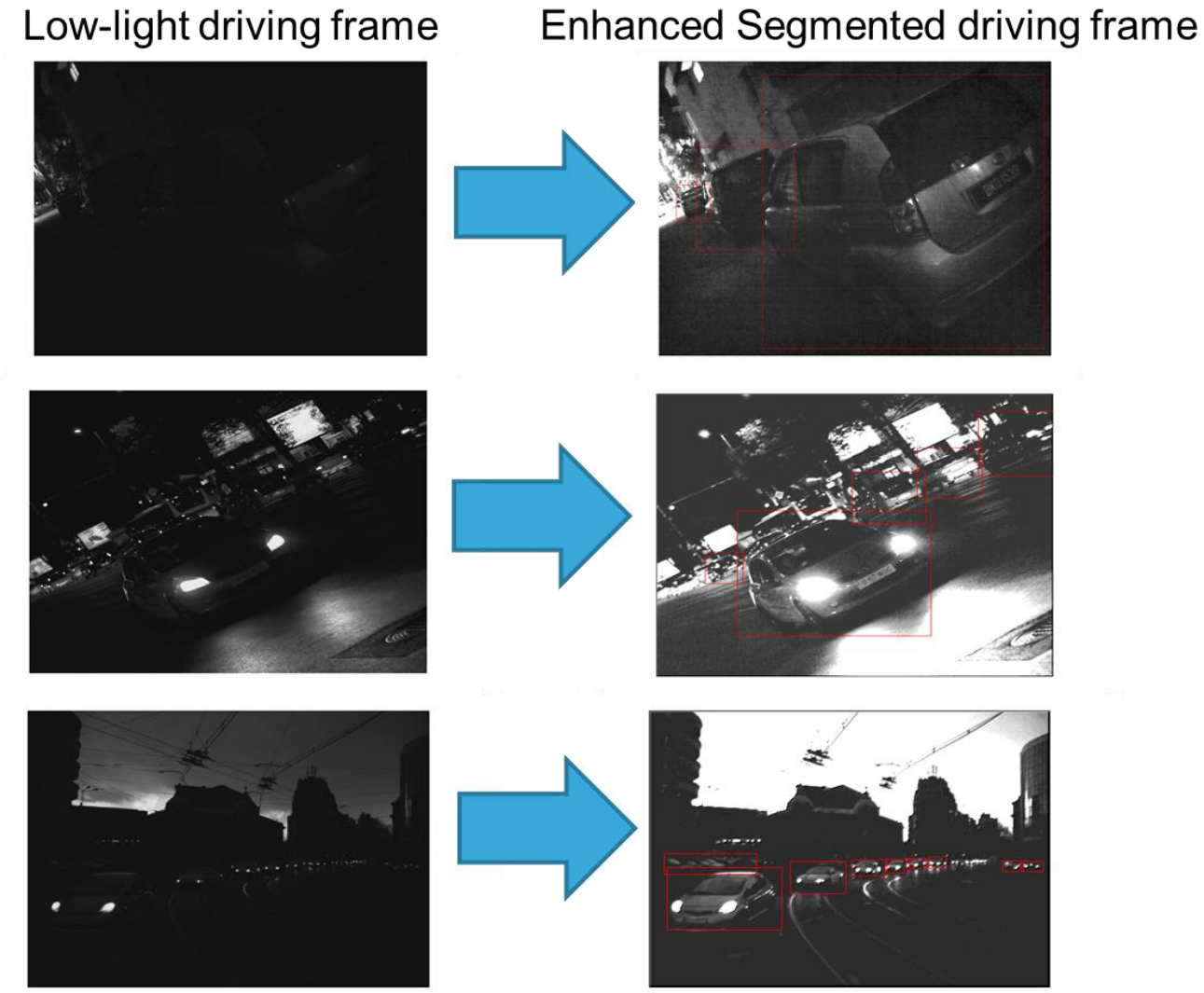


Figure 6. Some instances of the low-light driving frames with corresponding enhanced and segmented frames.

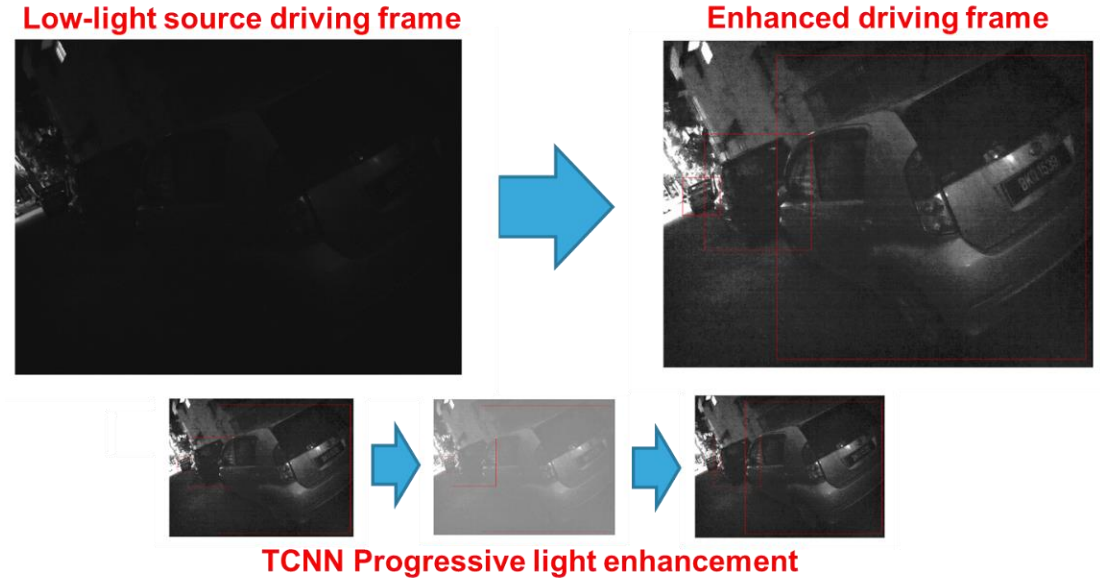


Figure 7. *Cont.*

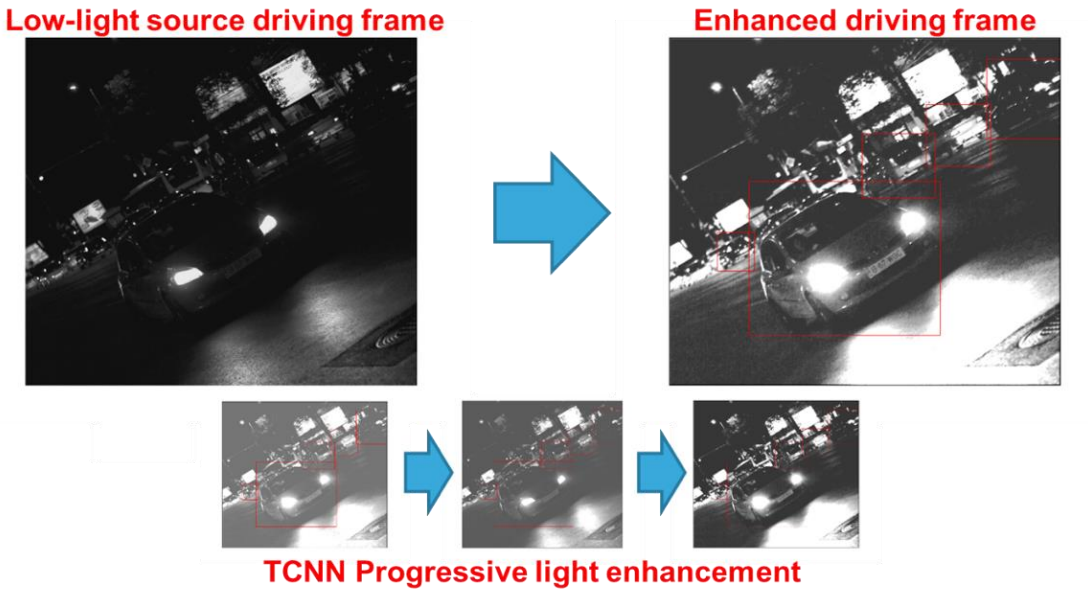


Figure 7. A detail of the enhanced low-light driving frames with embedding bounding-box (red box) and light-processing performed by the proposed TCNN as per Equations (5)–(7).

5. Discussion and Conclusions

The performance of our proposed pipeline to assess the safety risk of low-light driving scenarios has been confirmed by the experimental results described in the previous section. Compared with other methods in the literature, the implemented system has significant improvements that reveal its competitive advantage, such as the fact that there is no need of complex digital architecture or underlying hardware. The designed method overcomes the problems of the previous solutions because it uses only the TCNN block for enhancing low-light driving frames.

As introduced, TCNN can be implemented in U-VLSI analogic devices allowing real-time high computation and performance and low costs. A Sematic Segmentation Fully Convolutional Neural Network embedding Non-Local self-attention block is then used to assess the enhanced driving scene by a robust salient object detection and segmentation. The effective implemented pipeline is currently partially hosted over an embedded platform based on the STA1295 Accordo5 Silicon on Chip (SoC) platform produced by STMicroelectronics with a software environment which includes a distribution of YOCTO Linux O.S [28].

In order to improve the discriminating visual features and the ability to classify of the downstream classifier, the author is investigating the use of innovative self-attention mechanism as reported in scientific literature [33–36] as well the usage of a combined supervised/unsupervised approach embedding reinforcement learning and such use of hand-crafted features [20,37].

It is thought for the next evolutions of the proposed pipeline to make stronger the embedded method of domain adaptation by making use of hybrid deep approaches applied to stereoscopic input driving frames [38].

6. Patents

The reported information is covered by the following registered patents: IT Patent Nr. 102017000120714, 24 October 2017. IT Patent Nr. 102019000005868, 16 April 2018; IT Patent Nr. 102019000000133, 7 January 2019.

Funding: This research was funded by the National Funded Program 2014–2020 under grant agreement n. 1733, (ADAS + Project).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Heimberger, M.; Horgan, J.; Hughes, C.; McDonald, J.; Yogamani, S. Computer vision in automated parking systems: Design, implementation and challenges. *Image Vis. Comput.* **2017**, *68*, 88–101. [CrossRef]
2. Horgan, J.; Hughes, C.; McDonald, J.; Yogamani, S. Vision-Based Driver Assistance Systems: Survey, Taxonomy and Advances. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, New York, NY, USA, 15–18 September 2015; pp. 2032–2039.
3. Pham, L.H.; Tran, D.N.-N.; Jeon, J.W. Low-Light Image Enhancement for Autonomous Driving Systems using DriveRetinex-Net. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Seoul, Korea, 1–3 November 2020; pp. 1–5. [CrossRef]
4. Rundo, F.; Conoci, S.; Spampinato, C.; Leotta, R.; Trenta, F.; Battiato, S. Deep Neuro-Vision Embedded Architecture for Safety Assessment in Perceptive Advanced Driver Assistance Systems: The Pedestrian Tracking System Use-Case. *Front. Neuroinform.* **2021**, *15*, 667008. [CrossRef] [PubMed]
5. Rundo, F.; Petralia, S.; Fallica, G.; Conoci, S. A Nonlinear Pattern Recognition Pipeline for Ppg/Ecg Medical Assessments. In *Convegno Nazionale Sensori*; Sensors; Springer: Berlin/Heidelberg, Germany, 2018; pp. 473–480.
6. Trenta, F.; Conoci, S.; Rundo, F.; Battiato, S. Advanced Motion-Tracking System with Multi-Layers Deep Learning Framework for Innovative Car-Driver Drowsiness Monitoring. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–5.
7. Rundo, F.; Rinella, S.; Massimino, S.; Coco, M.; Fallica, G.; Parenti, R.; Conoci, S.; Perciavalle, V. An innovative deep learning algorithm for drowsiness detection from eeg signal. *Computation* **2019**, *7*, 13. [CrossRef]
8. Rundo, F.; Conoci, S.; Battiato, S.; Trenta, F.; Spampinato, C. Innovative Saliency Based Deep Driving Scene Understanding System for Automatic Safety Assessment in Next-Generation Cars. In Proceedings of the 2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE), Turin, Italy, 18–20 November 2020; pp. 1–6.
9. Rundo, F.; Spampinato, C.; Battiato, S.; Trenta, F.; Conoci, S. Advanced 1D Temporal Deep Dilated Convolutional Embedded Perceptual System for Fast Car-Driver Drowsiness Monitoring. In Proceedings of the 2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE), Turin, Italy, 18–20 November 2020; pp. 1–6.
10. Rundo, F.; Spampinato, C.; Conoci, S. Ad-hoc shallow neural network to learn hyper filtered photoplethysmographic (ppg) signal forefficient car-driver drowsiness monitoring. *Electronics* **2019**, *8*, 890. [CrossRef]
11. Guo, Y.; Lu, Y.; Liu, R.W.; Yang, M.; Chui, K.T. Low-Light Image Enhancement with Regularized Illumination Optimization and Deep Noise Suppression. *IEEE Access* **2020**, *8*, 145297–145315. [CrossRef]
12. Qu, Y.; Ou, Y.; Xiong, R. Low Illumination Enhancement for Object Detection In Self-Driving. In Proceedings of the 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), Dali, China, 6–8 December 2019; pp. 1738–1743. [CrossRef]
13. Chen, G.; Cao, H.; Conradt, J.; Tang, H.; Rohrbein, F.; Knoll, A. Event-Based Neuromorphic Vision for Autonomous Driving: A Paradigm Shift for Bio-Inspired Visual Sensing and Perception. *IEEE Signal Process. Mag.* **2020**, *37*, 34–49. [CrossRef]
14. Rashed, H.; Ramzy, M.; Vaquero, V.; El Sallab, A.; Sistu, G.; Yogamani, S. FuseMODNet: Real-Time Camera and LiDAR Based Moving Object Detection for Robust Low-Light Autonomous Driving. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 2393–2402. [CrossRef]
15. Deng, J.; Pang, G.; Wan, L.; Yu, Z. Low-light Image Enhancement based on Joint Decomposition and Denoising U-Net Network. In Proceedings of the 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Exeter, UK, 17–19 December 2020; pp. 883–888. [CrossRef]
16. Szankin, M.; Kwaśniewska, A.; Ruminski, J.; Nicolas, R. Road Condition Evaluation Using Fusion of Multiple Deep Models on Always-On Vision Processor. In Proceedings of the IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society, Washington, DC, USA, 21–23 October 2018; pp. 3273–3279. [CrossRef]
17. Yang, W.; Wang, W.; Huang, H.; Wang, S.; Liu, J. Sparse Gradient Regularized Deep Retinex Network for Robust Low-Light Image Enhancement. *IEEE Trans. Image Process.* **2021**, *30*, 2072–2086. [CrossRef] [PubMed]
18. Chua, L.O.; Yang, L. Cellular Neural Networks: Applications. *IEEE Trans. Circuits Syst.* **1988**, *35*, 1273–1290. [CrossRef]
19. Chua, L.O.; Yang, L. Cellular Neural Networks: Theory. *IEEE Trans. Circuits Syst.* **1988**, *35*, 1257–1272. [CrossRef]
20. Conoci, S.; Rundo, F.; Petralta, S.; Battiato, S. Advanced Skin Lesion Discrimination Pipeline for Early Melanoma Cancer Diagnosis towards PoC Devices. In Proceedings of the 2017 European Conference on Circuit Theory and Design (ECCTD), Catania, Italy, 4–6 September 2017; pp. 1–4. [CrossRef]
21. Mizutani, H. A New Learning Method for Multilayered Cellular Neural Networks. In Proceedings of the Third IEEE International Workshop on Cellular Neural Networks and their Applications (CNNA-94), Rome, Italy, 18–21 December 1994; pp. 195–200. [CrossRef]

22. Cardarilli, G.C.; Lojacono, R.; Salerno, M.; Sargeni, F. VLSI Implementation of a Cellular Neural Network with Programmable Control Operator. In Proceedings of the 36th Midwest Symposium on Circuits and Systems, Detroit, MI, USA, 16–18 August 1993; Volume 2, pp. 1089–1092. [CrossRef]
23. Roska, T.; Chua, L.O. Cellular Neural Networks with Nonlinear and Delay-Type Template Elements. In Proceedings of the IEEE International Workshop on Cellular Neural Networks and their Applications, Budapest, Hungary, 16–19 December 1990; pp. 12–25. [CrossRef]
24. Arena, P.; Baglio, S.; Fortuna, L.; Manganaro, G. Dynamics of State Controlled CNNs. In Proceedings of the 1996 IEEE International Symposium on Circuits and Systems. Circuits and Systems Connecting the World. ISCAS 96, Atlanta, GA, USA, 12–15 May 1996; Volume 3, pp. 56–59. [CrossRef]
25. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803. [CrossRef]
26. Rundo, F.; Banna, G.L.; Prezzavento, L.; Trenta, F.; Conoci, S.; Battiato, S. 3D Non-Local Neural Network: A Non-Invasive Biomarker for Immunotherapy Treatment Outcome Prediction. Case-Study: Metastatic Urothelial Carcinoma. *J. Imaging* **2020**, *6*, 133. [CrossRef] [PubMed]
27. Min, K.; Corso, J.J. TASED-Net: Temporally-Aggregating Spatial Encoder-Decoder Network for Video Saliency Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 2394–2403.
28. STMicroelectronics ACCORDO 5 Automotive Microcontroller. Available online: https://www.st.com/en/automotive-infotainment-and-telematics/automotive-infotainment-socs.html?icmp=tt4379_gl_pron_nov2016 (accessed on 2 July 2019).
29. Rundo, F.; Leotta, R.; Battiato, S. Real-Time Deep Neuro-Vision Embedded Processing System for Saliency-based Car Driving Safety Monitoring. In Proceedings of the 2021 4th International Conference on Circuits, Systems and Simulation (ICCSS), Kuala Lumpur, Malaysia, 26–28 May 2021; pp. 218–224. [CrossRef]
30. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 year, 1000 km: The oxford robotcar dataset. *Int. J. Robot. Res.* **2016**, *36*, 0278364916679498. [CrossRef]
31. Loh, Y.P.; Chan, C.S. Getting to know low-light images with the Exclusively Dark dataset. *Comput. Vis. Image Underst.* **2019**, *178*, 30–42. [CrossRef]
32. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A Survey of Deep Learning-based Object Detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
33. Jian, M.; Lam, K.; Dong, J.; Shen, L. Visual-Patch-Attention-Aware Saliency Detection. *IEEE Trans. Cybern.* **2015**, *45*, 1575–1586. [CrossRef] [PubMed]
34. Jian, M.; Zhang, W.; Yu, H.; Cui, C.; Nie, X.; Zhang, H.; Yin, Y. Saliency detection based on directional patches extraction and principal local color contrast. *J. Vis. Commun. Image Represent.* **2018**, *57*, 1–11. [CrossRef]
35. Jian, M.; Wang, J.; Yu, H.; Wang, G.; Meng, X.; Yang, L.; Dong, J.; Yin, Y. Visual saliency detection by integrating spatial position prior of object with background cues. *Expert Syst. Appl.* **2021**, *168*, 114219. [CrossRef]
36. Jian, M.; Qi, Q.; Dong, J.; Yin, Y.; Lam, K.-M. Integrating QDWD with pattern distinctness and local contrast for underwater saliency detection. *J. Vis. Commun. Image Represent.* **2018**, *53*, 31–41. [CrossRef]
37. Rundo, F. Deep LSTM with Reinforcement Learning Layer for Financial Trend Prediction in FX High Frequency Trading Systems. *Appl. Sci.* **2019**, *9*, 4460. [CrossRef]
38. Ortis, A.; Rundo, F.; Di Giore, G.; Battiato, S. Adaptive Compression of Stereoscopic Images. In *Image Analysis and Processing–ICIAP 2013*; Petrosino, A., Ed.; ICIAP 2013; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8156. [CrossRef]