electronics

MDPI

Article

Channel-Attention-Enhanced LSTM Neural Network Decoder and Equalizer for RSE-Based Optical Camera Communications

Peng Ling, Maolin Li and Weipeng Guan *

School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China; 201930133152@mail.scut.edu.cn (P.L.); 202030461169@mail.scut.edu.cn (M.L.)
* Correspondence: augwpscut@mail.scut.edu.cn

Abstract: In an RGB-LED-based optical camera communication system, it is an essential goal to have better performance in the data rate and BER. However, in a higher symbol rate, due to the conventional sampling algorithm, the deterioration of transmission performance brought by the inter-symbol interference and inter-channel interference is significant. Innovatively, in this paper, the sub-image obtained by a captured frame of received video is encoded by a channel-attention-Net-based encoder to generate a descriptor without existing sampling methods. Moreover, we propose an LSTM-based equalizer to decode the descriptor and mitigate transmission performance deterioration. Utilizing the long-short-term memory of an LSTM unit, an equalizer not only can reduce bit error rates but also increase the data rate. The experimental results show that at a symbol rate of 46 kbaud/s, a record-high data rate at 44.03 kbit/s is achieved under random data transmission while still meeting the pre-forward error correction requirement.

Keywords: optical camera communication; rolling shutter effect; visible light communication; machine learning; machine vision

check for updates

Citation: Ling, P.; Li, M.; Guan, W. Channel-Attention-Enhanced LSTM Neural Network Decoder and Equalizer for RSE-Based Optical Camera Communications. *Electronics* **2022**, *11*, 1272. https://doi.org/10.3390/electronics11081272

Academic Editor: Paulo Monteiro

Received: 23 March 2022
Accepted: 15 April 2022
Published: 17 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recently, visible light communication (VLC) has attracted considerable attention as a short-range communication technology due to its characteristics of higher bandwidth and lower time delay [1]; it is even regarded as one of the next wireless communication technologies (6G) [2]. VLC technology provides opportunities to the applications of indoor positioning, mobile payment and navigation [3]. In most research, VLC needs photodiodes as dedicated receivers, which delays its commercialization. Thus, optical camera communication (OCC) mostly using a charge-coupled device (CCD) or complementary-metal-oxide-semiconductor (CMOS) cameras as the receiver is researched as a specific type of VLC to overcome the issue [4]. CMOS-based commercial optical cameras are being widely utilized in mobile phones with their rapid development; thus, it provides conditions for OCC technology development [5].

The color-shift keying (CSK) modulation scheme recommended by [6] has been widely adopted in RGB-LED-based OCC systems, offering higher communication efficiency. Three bits are represented by one symbol in an 8-CSK modulation scheme [7], which potentially enhances the data rate performance compared with the on–off keying modulation scheme [8]. However, the inter-symbol interference (ISI) will be introduced by the CMOS-based camera's sampling [1], which will significantly degrade the RGB-LED-based OCC system performance. Most conventional decoding algorithms employed for mitigating ISI and inter-channel interference (ICI) [9] do not perform well in higher-speed transmission environments and only offer data rates of 8.64 kbit/s and 17 kbit/s, respectively [10,11]. More ingenious algorithms are desperately needed for making innovations for conventional sampling and equalization methods.

In recent years, deep learning technology represented by convolutional neural network (CNN) and recurrent neural network (RNN) provides methods to solve these problems

faced by decoding in OCC systems [12]. In the deep learning field, CNN-based neural networks are sought highly for solving the image classification and region-of-interest (ROI) detection problems, since they specialize in features extraction and nonlinear mapping. However, most previous research only focus on using CNNs for ROI detection [13,14] and equalization [15], which means that the processing of demodulation still requires signal sampling. Conventional signal sampling algorithms are susceptible to ISI and ICI, resulting in performance deterioration. Introducing CNN-based decoding algorithms for extracting features for images received provides a potential solution for avoiding signal sampling and mitigates ISI and ICI. Toward the time-sequence prediction problem, RNN seems to be a more feasible solution. Long-short-term-memory (LSTM)-based RNN [16] is widely used in natural language processing (NLP) [17], emotional analysis [18], and other time-sequence problems. In VLC, an LSTM-based monitor is used to estimate network performance by calculating the signal-to-noise ratio (OSNR) [19]. However, in the field of rolling-shutter-effect (RSE)-based OCC, the application of RNN is still in its infancy, no researchers pay attention to utilizing LSTM-based equalizers, which reflects that most of the existing equalizers only take spatial-wise equalization into consideration.

Combining their advantages of feature extraction and learning to store information over time intervals, we provide an appropriate decoding scheme fusion of the function of demodulation and spatial-wise and time-wise equalization. In this work, to our knowledge, we are the first to put forward a CNN-LSTM-based decoding scheme in which CNN serves as an encoder, and LSTM serves as a decoder and equalizer in OCC. The main contributions of this work are summarized as follows:

- We design a novel neural network named channel-attention-enhanced LSTM (CAE-LSTM-Net) for an OCC decoding system, which is inspired by the channel attention mechanism and LSTM. It is an end-to-end network that can decode signals without conventional sampling from images received by a CMOS-based camera, which is different from all existing algorithms for RGB-LED-based RSE-based OCC systems. Since our decoding method does not need conventional sampling, it is not easily affected by sampling offset caused by ISI. Based on our equalizer, we also fuse the function of spatial-wise and time-wise equalization that is for the first time considered by an equalizer for an OCC system.
- We propose a header-location-based image segmentation algorithm with satisfactory supportability for precise sub-image demodulation that combines signal tracing processing and gamma correction.
- To our knowledge, the performance of the proposed decoding algorithm in data rate and BER is record-high in the RSE-based OCC. We also prove experimentally that the decoding scheme based on CAE-LSTM-Net outperforms those comprised of other existing CNNs when the transmitted symbol is modulated by 8-CSK.

Organization

The remainder of this paper is organized as follows. Section 2 introduces the background. Section 3 explains the methodology of our CAE-LSTM-Net-based decoding algorithm. Section 4 introduces the experimental setup. Section 5 presents the experimental evaluation and analysis. Finally, in Section 6, the conclusion is presented.

2. Background

2.1. Principal of Using CMOS Sensor in OCC

An image sensor is classified by two main technologies: one is CCD supporting global shutter, and the other is CMOS supporting rolling shutter, as we can learn from [20–22]. The construction process of an image using the global shutter effect is shown in Figure 1. All the pixels on the sensor are exposed simultaneously, and the readout data are accessed at the end of exposure. With an RSE-based camera, the captured image is generated row by row of pixels, which means that the pixels exposure and data readout are performed sequentially by each row in a CMOS sensor. All scanlines captured at different exposure

times are stitched into an image at last, as shown in Figure 2. Utilizing an RSE-based camera, changing states of light source will be captured in one image. However, if the transmitter, comprised of RGB-LED in our work, is modulated at a higher-speed, ISI will be introduced probably. The main incentive of ISI is that there are overlaps between the exposure time of each row, and two states of light source will be captured in them, which will result in an invalidation of pixel-row per symbol in a large proportion of rows. Additionally, there is a processing gap time between the constructions of two adjacent frames when no row can be activated during this period. In [23,24], a special packet design and packet combination are proposed to solve the problem in which the signal emitted during the processing gap-time of the CMOS sensor cannot be captured. However, finding a more effective scheme to mitigate intense ISI in higher-speed signal transmission is still an impediment for OCC technology development.

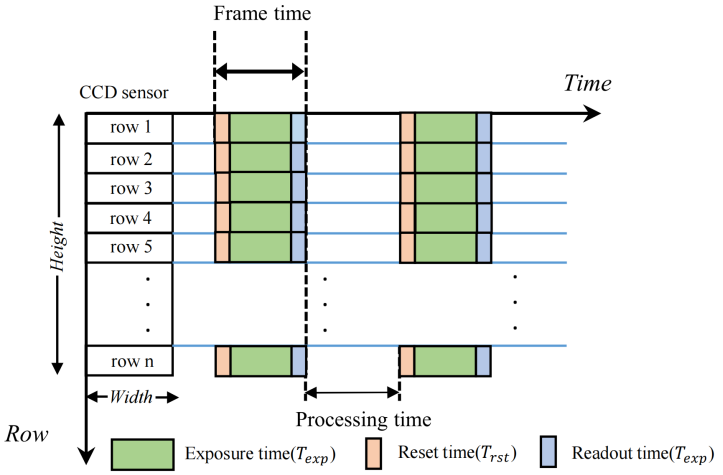

Figure 1. Sketch map of the global shutter of the CCD sensor.

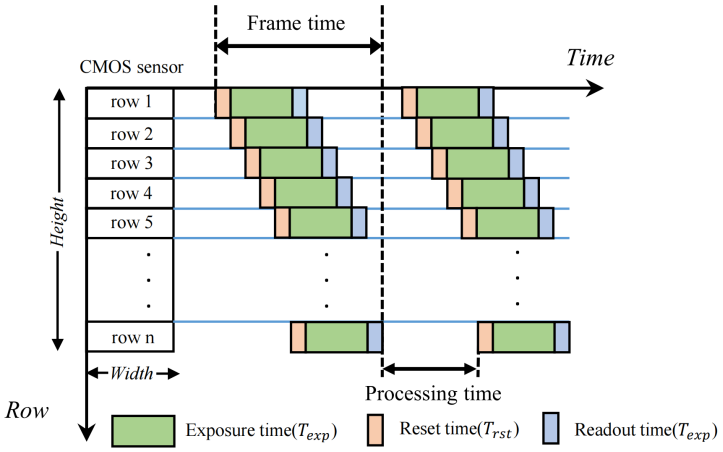

Figure 2. Sketch map of the rolling shutter of the CMOS sensor.

2.2. Principal of CNN-Based Encoder in OCC

CNN has been used in image recognition for identifying humans, objects, etc. For the task of feature extraction, CNNs utilizes convolution to generate output information, which is known as a "Feature map". The convolution kernel which contains the transformation arguments that are applied to the input argument can easily extract the features that it wants to highlight. After convolution, nonlinear activation function and pooling layers are extensively used in CNN to explicitly model the nonlinear relationship between input and output information. In the training process, an optimization algorithm is used to iterate the optimal parameters of kernels that characterize the network perception. In the specific case of image classification, softmax operation in the last layer is responsible for

generating the image class and its corresponding probability value. Essentially, a great majority of CNNs that applied in the classification task are comprised of two stages: one is the convolution-based feature extraction stage, and the other is the ANN-like nonlinear mapping stage, which is used to transform a features descriptor generated by the first stage into an output classification result, as shown in Figure 3. However, the output result is bounded, and it means that all layers in CNN are only activated by current input information, which is highly effective for an independent classification problem but not for time-variant application scenarios. Therefore, in this work, we take advantage of CNN as a feature encoder to generate a descriptor of input image, and it is not configured to accomplish the task of transforming the higher-level features to an output class.

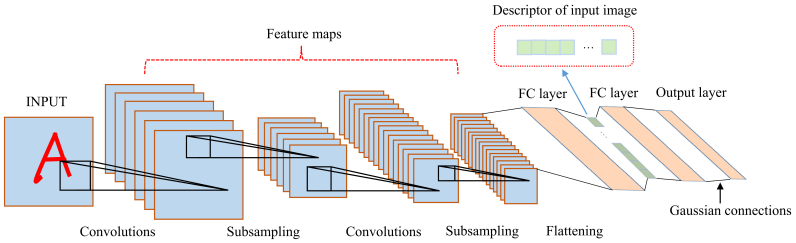

Figure 3. Sketch map of the feature maps obtained by LeNet-5 [25]. The descriptor of input image is generated in the second stage of neural network.

2.3. The LSTM Model

Recurrent neural networks (RNNs) are fundamentally different from the feed-forward neural network. In RNN, the temporal correlations between previous information and current circumstance is modeled explicitly, which is aiming at solving sequence-based problems. However, it is difficult to learn long-range dependencies with traditional RNNs due to the gradient vanishing or exploding problem [26,27]. In order to overcome these issues, in [16], Hochreiter et al. propose the long short-term memory (LSTM) architecture that is further improved by Gers et al. [28]. The structure of the LSTM unit and the sequence classification model are depicted in Figures 4 and 5. In brief, a gating mechanism is introduced by their works, which provides avenues for increasing network insensitivity to gap length. The formulations of all function nodes in an LSTM unit are given by:

$$f_t = \sigma\Big(W_{fx}x_t + W_{fh}h_{t-1} + b_f\Big), \tag{1}$$

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \tag{2}$$

$$\widetilde{C}_t = \phi\big(W_{\widetilde{C}x}x_t + W_{\widetilde{C}h}h_{t-1} + b_g\big), \tag{3}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o), \tag{4}$$

$$C_t = \widetilde{C}_t \odot i_t + C_{t-1} \odot f_t, \tag{5}$$

$$h_t = \phi(C_t) \odot o_t. \tag{6}$$

where W_{fx}, W_{fh}, W_{ix}, W_{ih}, $W_{\widetilde{C}x}$, $W_{\widetilde{C}h}$, W_{ox}, and W_{oh} are weight matrices corresponding to the input of activation functions; σ and ϕ represent the sigmoid activation function and tanh function respectively; and $\odot$ means an element-wise multiplication. In the deep learning literature, there are three gates in an LSTM block, the forget gate f, the input gate i, and the output gate o, and all of them take advantage of the sigmoid output range from 0 to 1. It is intuitive that the decisions for the three gates are dependent on the previous output h_{t-1} and the current input x_t. Specifically, the input gate, forget gate, and output gate are responsible for deciding what to preserve in the internal state, what to forget from the previous state, and which input signals should pass from $\widetilde{C}t$ to the output h_t, respectively. In application, the dimension of input signal x_t should be specified at first as a hyperparameter, and it depends on the dimension of all internal vectors. The relationship between input x_t and output h_t is associated with the memory cell state

C_{t-1} and intermediate output h_{t-1} at the $t-1$th time step. It means that utilizing the LSTM block, we can construct transformation from the previous input and current input to the current output, and we can realize equalization to mitigate ISI with consideration of information brought by all kinds of symbol arrangement. Such equalization contributes to both BER reduction and data rate increasing, taking advantage of long-short-term memory.

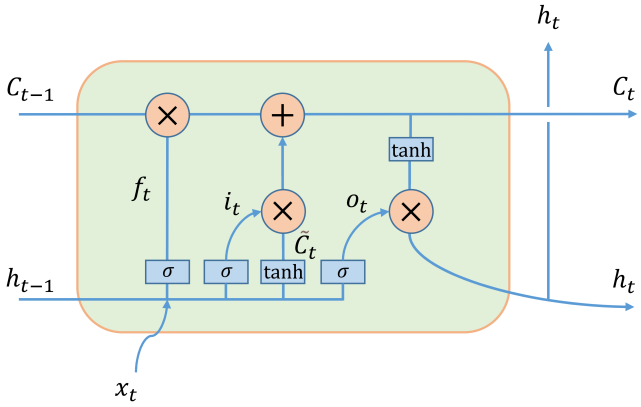

Figure 4. Sketch map of the LSTM unit.

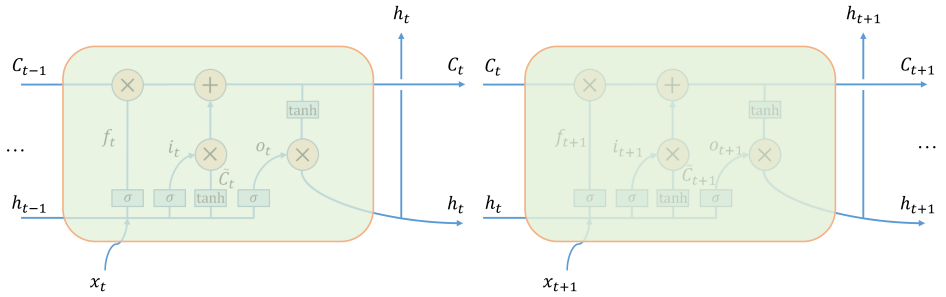

Figure 5. The sequence classification model based on LSTM unit. x_t is the input at time step t and h_t is the output at time step t.

3. Proposed Neural-Network-Based Decoding Scheme

3.1. Proposed Neural Network

This paper propose a neural network named CAE-LSTM-Net whose structure is shown in Figure 6. The CAE-LSTM-Net comprises of two parts of neural network: one is channel-attention-Net serving as an encoder, and the other is LSTM unit serving as a decoder and equalizer. We introduce this network with the goal of improving the quality of representations and equalization produced by the feature encoder and equalizer. Inspired by channel attention [29], we take into consideration channel-wise feature recalibration in CNN-based descriptor design. As shown in the middle part of Figure 6, the descriptor of the sub-image from the output of the encoder at each time step is a one-dimensional vector generated by flattening the feature maps after channel-wise recalibration. Meanwhile, for the purpose of mitigating the ISI introduced by the exposure time overlap of the previous symbol shown below in Figure 6 in sub-image demodulation, and enhancing the accuracy of recognizing a specific data sequence (i.e., header of each packet) in the data packet, it occurs to us that we can take advantage of the correlations between previous information and current circumstance learned by the LSTM unit and build an equalizer to learn different nonlinear equalization mapping under different symbols arrangement. For instance, as shown in Figure 6, at the second time step, the sub-image is significantly distorted due to ISI brought by the previous symbol at the first time step, which may cause an error decoded bit. The proposed LSTM-based equalizer is capable of learning an appropriate pattern that avoids decoding the current sub-image into the previous one bit, using its short-term memory.

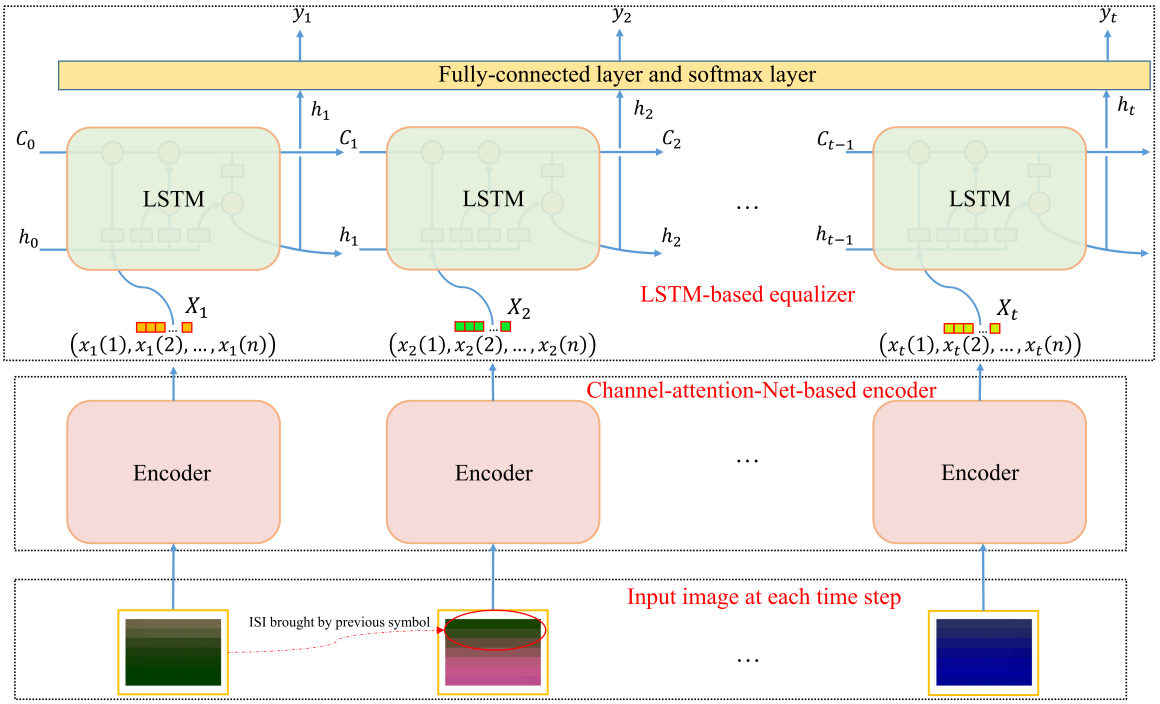

Figure 6. Sketch map of the structure of the proposed CAE-LSTM-Net.

3.2. Channel-Attention-Net-Based Encoder of Proposed Neural Network

How to generate an appropriate descriptor of a sub-image using its internal feature is still a challenge. Theoretically, there are two aspects of CNN design that deserve special attention: one is better modeling spatial dependencies [30,31], and the other is interdependencies between channels. For the task of extracting strong discriminative features, we enhance the network's sensitivity of channel-wise and plane-wise informative features by using the channel attention mechanism and residual mapping. As shown in Figure 7, the proposed encoder consists of two main architectures: residual unit and recalibration unit. These two architectures are shown in Figure 8. In basic blocks, the residual unit is responsible for fitting spatial residual distribution, and it benefits CNN to learn constructive solutions in backpropagation, resulting in a reduction of training error [32]. To transform lower-level extracted features to higher-level, we lay out two multichannel kernels in each residual unit, adopting a ReLU layer to model the nonlinear relationship between each layer of the feature map. A normal residual structure is not quite sensing the features in the third dimension ideally. Therefore, we design a recalibration unit to explicitly model channel interdependencies, using channel-wise global information to recalibrate filter responses. Utilizing the global layer and fully connected layer, we obtain the characterization of each channel of the input feature map. BatchNormalization layers constrain the output vector that is in the same distribution, which improves the smoothness of the optimization landscape in the process of iterating out the weight of each channel of input feature maps [33]. In the last stage of the encoder, the generation of a vector descriptor of the input sub-image is achieved by using average pooling. In details of the channel-attention-Net-based encoder, the size of the input image is 7×10, and the sizes of all convolution kernels are set to 3×3.

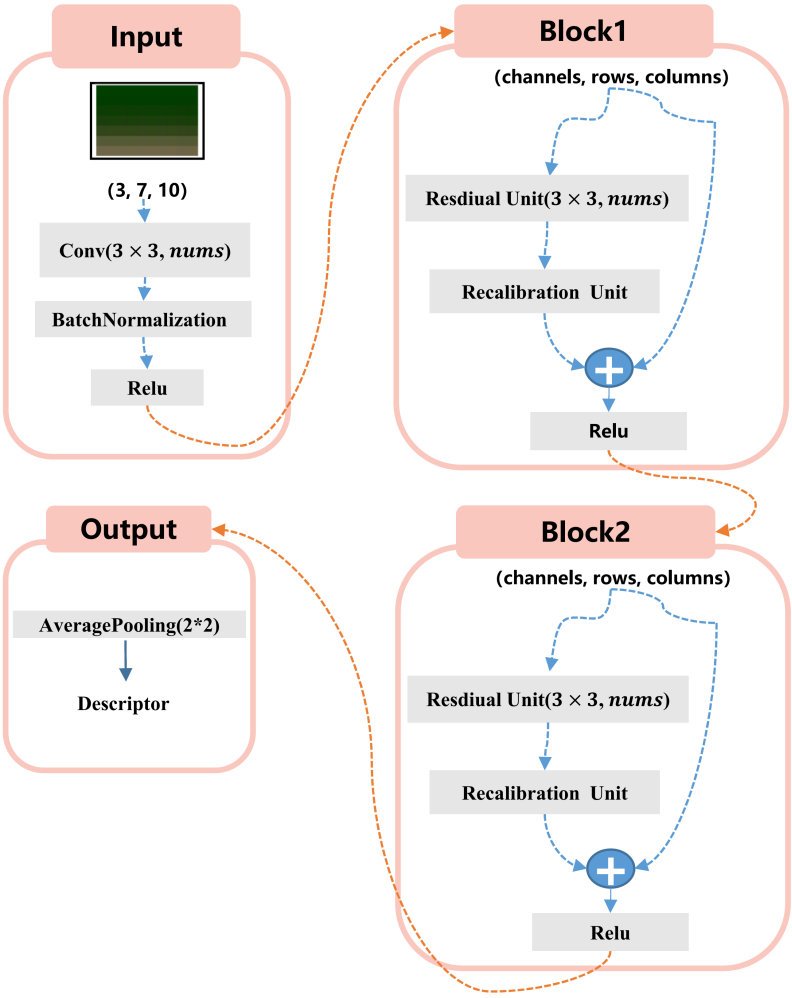

Figure 7. The structure of channel-attention-Net-based encoder which comprised of two basic blocks in this case.

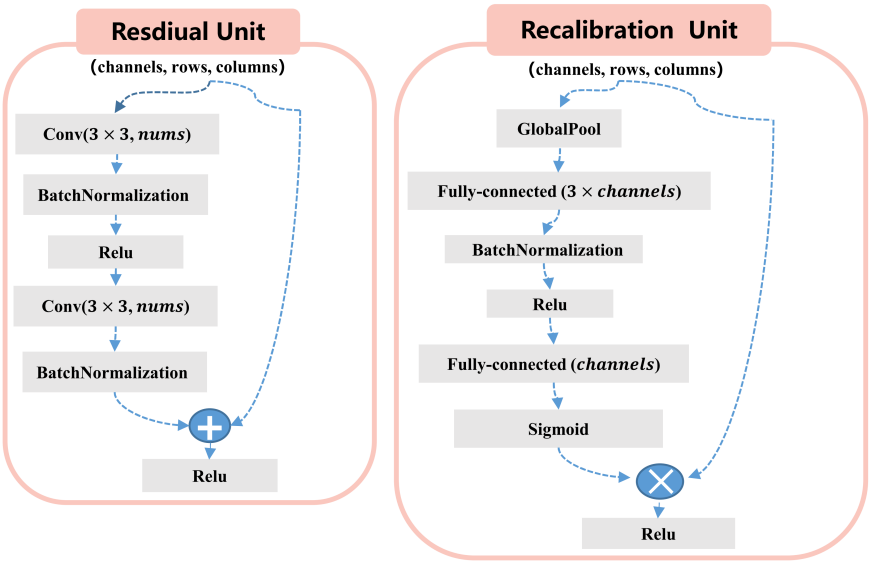

Figure 8. Sketch map of the architectures of residual unit and recalibration unit in basic block.

3.3. LSTM-Based Equalizer of Proposed Neural Network

The proposed equalizer based on LSTM is shown in Figure 6, which comprised an vector input layer, a hidden layer, and a classification layer. In this work, LSTM is

introduced to decode a vector descriptor generated by the proposed encoder and equalize channel distortion, which is equal to a time-series classification problem. In the "one-to-one" form, LSTM takes an input, a time-series window t of descriptor vector x_t (one descriptor vector x at each time step of window t), and at each time step of the time-series window t, it finally outputs a classification result, which is given by the softmax vector.

The overlaps between the exposure time of symbols cause a degree of ISI happening in the sub-image that needs to be demodulated now (at time step t). The generation of an output vector of the LSTM unit at time step t is a comprehensive consideration of long-short-term memory and the latest training data vector, which means that parameters in short-term memory can be used to quickly process ISI brought by a previous one symbol, and those parameters in long-term memory are used to increase the accuracy of recognizing symbols of specific data transmitted repeatedly, which contributes to recognizing more headers of data packets so that more packets could be decoded. Due to the effective use of long-short-term memory by an LSTM unit, it will learn a pattern so that it can make specific compensation for mitigating ISI happening in the current sub-image decoding after obtaining the output vector of the previous time steps so as to reduce the classification error.

3.4. Decoding Scheme

Our proposed decoding scheme can be summarized briefly: we segment each extracted frame into a series of sub-images in order and decode one sub-image in one time step using CAE-LSTM-Net. The proposed decoding scheme of the RGB-LED-based OCC system is shown in Figure 9.

In this OCC system, each data packet is transmitted three times repeatedly within (1/frame rate) second to ensure that a data packet can be recorded completely at least once in a captured frame. Each packet transmission had a 16-bit header and a payload, as shown in Figure 10. The signal is transmitted by a video recording with 60 frames per second (fps) and decoded offline. In an optical camera, the resolution of each frame in a captured video is 1080×1920. Firstly, an appropriate column of pixels is selected, and image enhancement is performed in these pixels. Utilizing the averaging-based signal tracing algorithm proposed in [9] and gamma correction, the signal intensity can be normalized, and the nonlinear distortion introduced by hardware can be compensated. These processes also make a large contribution to increasing the gaps between symbols and improving the recognition rate. The 8-CSK modulation is adopted in this work; it means that on–off keying is employed for each channel so that each symbol includes three bits. Before image incision, the reference position should be sought out by the threshold method [34], which is called header location hereinafter. Since the header of each packet transmission only comprises of the two symbols, one is the "brightness", and the other is the "darkness", both of them are easily observed. As shown in Figure 9a–f, a G channel signal best serves to recognize the header symbol. However, even though in the demodulation of the most distinct symbols, we can find that intense distortion happens in the other two channels, which means that the performance of conventional decoding algorithms with a sampling process will be deteriorated in most other cases as well. It is the reason why conventional algorithms are probably not feasible solutions for higher-rate transmission. After header location and interpolation, the sub-image with a size of 9 pixels in row is cut starting from the reference position. In the second dimension, if L is the column of selected pixels, then the column matrix of a sub-image is set from $L - 4$ to $L + 5$. A frame is regarded as a basic processing sequence; we input all sub-images to the trained channel-attention-enhanced LSTM sequentially from time step 0 to time step t. The output sequence comprised the output from time step 0 to time step t in chronological order. Each frame is used to recover only one data packet in timing recovery.

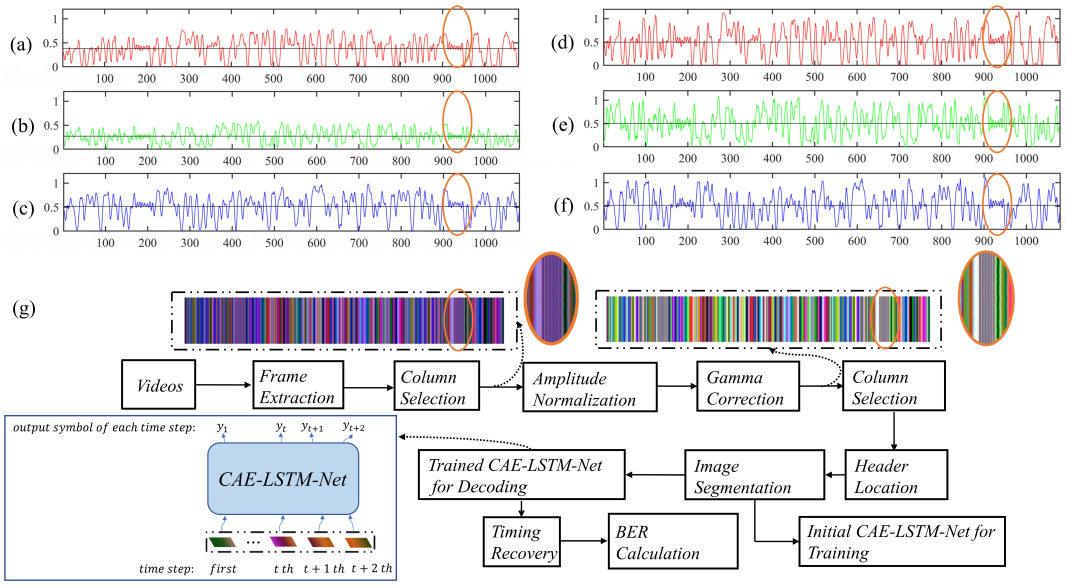

Figure 9. Offline processing of OCC decoding scheme. (**a**–**c**) are respectively R, G, and B signals of the raw column selected; (**d**–**f**) are respectively the R, G, and B signals of the selected column after enhancement; (**g**) Block diagram of the proposed offline OCC decoding scheme.

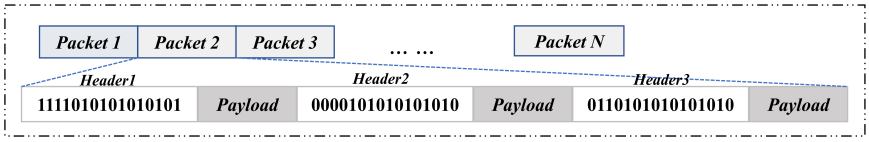

Figure 10. The signal construction with repetitive packet transmission in a color channel. For all packets, the headers comprised the same three sequences.

4. Experimental Setup

4.1. Dataset Setup

The experimental setup is shown in Figure 11. The transmitted packet constructed from random data was fed into a field-programmable gate array (FPGA, Xilinx Spartan 6, XC6SLX16). The symbol rate was set from 32 kbaud/s to 46 kbaud/s. We used the output signal from I/O port to drive an RGB-LED. A mobile phone (HUAWEI P20 PRO) with a CMOS-based optical camera was set at the data receiver side after a 40-CM free-space signal transmission. Note that a convex-convex lens and a diffuser were set in front of the phone. For the experiment, the mobile phone records a video at a frame rate of 60 fps, and the resolution is 1080 × 1920. For the parameters of an optical camera, the sensitivity (ISO) was set to 250, and the other parameters were set automatically by the software.

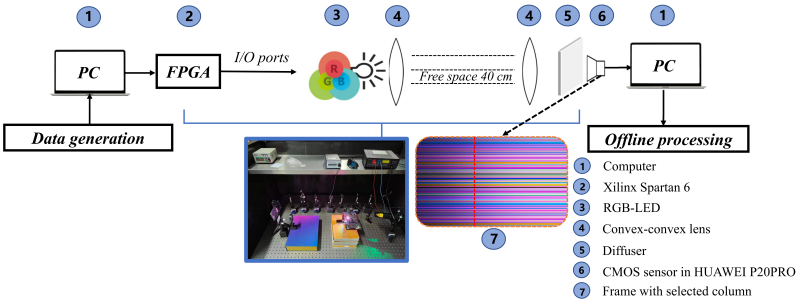

Figure 11. Block diagram and scene of the experimental setup. CMOS-based camera in HUAWEI P20PRO is the receiver and single RGB-LED drived by I/O ports of FPGA is the transmitter.

4.2. Training Detail

For a traditional equalizer, the weights are reset after each retraining, so there is no memory effect. However, neural networks have a memory effect on the training

samples, which is harmful to communication systems, because obviously, the sequence to be transmitted should be unpredictable. Thus, in this work, we used two videos capturing random messages generated separately for the training process and test process in order to avoid the performance over estimation. The number of training epochs is set to 25 and the Adaptive Momentum Estimation (Adam) optimizer is used after calculating the Softmax loss. Note that some hyperparameters such as regularization items, the number of basic blocks and kernels, and the number of hidden neurons in LSTM will be changed under different symbol rate conditions, for the purpose of searching global optimal solution effectively.

5. Results and Discussion

5.1. Sensitivity of Parameters

The sensitivity of neural network performance to some main structure parameters is investigated in this study. Table 1 summarizes the optimal parameters of our neural network. In this table, Kernels, Blocks, Hidden numbers, and LSTM layers mean the quantity of kernels in each convolutional layer, blocks in the channel-attention-Net-based encoder, hidden neurons in the LSTM unit, and LSTM layers, respectively. Figure 12a demonstrates the BER performance at various data rates. At lower-rate scenarios, we can see that the enhancement brought by more blocks is not significant. However, when the signal is modulated in a higher-rate at 40 to 46 kbaud/s, more blocks in the proposed encoder are necessary to generate a higher-level descriptor. Additionally, more convolutional kernels are conducive to improve the identification ability of the convolutional layer to the input map, as shown in Figure 12c, since there will be more feature extractors. Furthermore, it can be concluded from Figure 12b that there will be a more desirable performance when increasing the complexity of the LSTM-based equalizer, but we should consider that it can compromise the computational complexity and system performance.

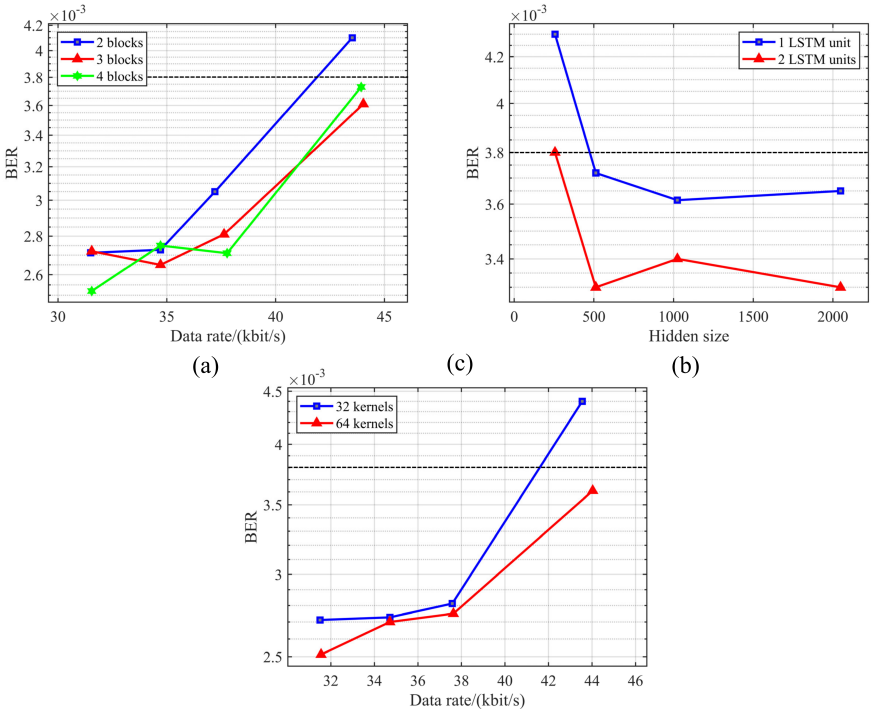

Figure 12. BER performance under different parameters of neural network structures. In addition to the parameters investigated in each experiment, the other parameters are set as shown in Table 1. (**a**) Sensitivity of neural network performance to the quantity of blocks in encoder; (**b**) Under 46 kbaud/s, sensitivity of neural network performance to the quantity of hidden neurons and layers of LSTM-based equalizer; (**c**) Sensitivity of neural network performance to the quantity of convolutional kernels in encoder.

Table 1. Optimal parameters in compromising consideration.

Symbol Rate (kbaud/s)	Kernels	Blocks	Hidden Numbers	LSTM Layers
32	32	2	512	1
36	32	2	512	1
40	32	3	512	1
46	64	3	1024	1

5.2. Performance Comparison

Table 2 summarizes the performance comparisons in data rate and BER of our RGB-based OCC system with the proposed decoding algorithms and several recently proposed OCC systems. In RSE-based OCC systems, to increase the transmission distance and data rate, most researchers make an effort to modify the modulation and demodulation algorithms and replace white light LED with RGB-LED. In [8,35], researchers adopt nonlinear fitting algorithms to equalize the distortion brought by the sampling offset, and they make progress in the performance of data rate and BER of the OCC system. In [36], this study utilizes ANN as an amplitude equalizer, but there is no significant improvement of data rate and BER performance. As the symbol rate goes faster, the performance deterioration caused by intensive ISI and ICI is inevitable in higher-speed OCC, which is mainly due to the sampling offset, which means that we need a much better decoding scheme and equalizer. In [5], this study proposes a Manchester-code-based decoding algorithm without sampling in mobile OCC, using CNN. However, owing to the disadvantages of Manchester code in utilizing RGB mode and the limited ability of the proposed CNN, it only adopts a single LED and has a data rate of about 2 kbit/s. Compared with the algorithm proposed in [15], in a similar experimental environment, under the circumstance without using an XOR compensation scheme, we achieve a data rate growth from 39 to 44.03 kbit/s. Actually, XOR compensation is not very meaningful to an OCC system, since it decodes much redundant data. In [15], 2D-CNN is just used for equalization after sampling by conventional algorithm, which means that it can be replaced by an ANN-based or SVM-based equalizer, and it is vulnerable to sampling offset caused by ISI. Giving the credit to higher-quality feature extraction and well-designed equalizer, the CAE-LSTM-Net proposed in this paper is not susceptible to interference, and it achieves a record-high data rate and the FEC requirement, combining feature extraction and equalization in an end-to-end way without sampling for decoding. Frankly speaking, our algorithm needs more computing resources because of its higher-complexity neural network, but it probably can be solved by special chips and circuits.

Table 2. Performance comparison.

Hardware	Technics	Data Rate	BER	Distance	Time
RSE-based CMOS sensor	Manchester decoding, Block detection [23]	1.1 kbit/s	No mention	35 cm	2012
	OOK(Second-order polynomial fitting) [8]	0.896 kbit/s	3.25×10^{-1} to 6.05×10^{-4}	25 cm	2015
	OOK(Polynomial fitting) [35]	1.68 kbit/s	3×10^{-3} to 4×10^{-6}	25 cm	2015
	CSK(RGB,MIMO) [37]	2.88 kbit/s	3.16×10^{-3}	10 cm	2016
	OOK(UFSOOK) [24]	10.32 kbit/s	1.01×10^{-4}	20cm	2017
	OOK(LR) [38]	0.78 kbit/s	Less than 3.8×10^{-3}	1.5 m	2019
	CSK(RGB channel separation) [39]	0.96 kbit/s	Less than 3.8×10^{-3}	2.5 m	2020
	V4PPM [40]	4.8 kbit/s	No mention	4 m	2015
	OOK and FSK(RGB-MIMO,ANN) [36]	1.2 kbit/s	3.53×10^{-3}	2.5 m	2021
	OOK (XOR compensation and 2D-CNN equalizer) [15]	43.7 kbit/s	3.80×10^{-3}	40 cm	2019
	Manchester code and CNN-based decoder and equalizer [5]	2.16 kbit/s	3.80×10^{-5}	45 cm	2020
	This work	44.03 kbit/s	3.61×10^{-3}	40 cm	2022

5.3. Ablation Study

In order to demonstrate that the proposed CAE-LSTM-Net-based decoding algorithm has a state-of-the-art performance, we experimentally investigate the performance improvement brought by the LSTM-based equalizer and channel-attention-Net-based encoder. We first perform experimental comparison at two different conditions: one is with an LSTM-based equalizer, and the other is without an LSTM-based equalizer. In the second condition, we place the Softmax layer following the channel-attention-Net-based encoder, and we use its output as the classification result at time step t. Additionally, the traditional equalizer based on ANN is also investigated for comparison. Figures 13 and 14 illustrate the comparison of BER performance and data rate performance at a symbol rate from 32 to 46 kbaud/s. The BER performance of the scheme with a channel-attention-Net-based encoder and ANN-based equalizer is better than that without an ANN-based equalizer, but the difference between them is not apparent. The reason is that each classification problem is separate, which means that a sharp accuracy deterioration will happen when most of the pixels in the sub-image stand for the wrong symbol due to ISI. Using the proposed CAE-LSTM-Net, the short-term memory of LSTM is appropriately utilized to mitigate decoding system performance deterioration. As shown in Figure 13, using CAE-LSTM-Net, the BER can be reduced from 9.23×10^{-3} to 3.61×10^{-3} at a symbol rate of 46 kbaud/s, and BERs are meeting the pre-forward error correction (FEC, $BER = 3.8 \times 10^{-3}$) requirement in all cases. As shown in Figure 14, the proposed LSTM-based equalizer outperforms those with an ANN-based equalizer or without a specific equalizer in data rate performance. The data rate is significantly improved by using an LSTM-based equalizer. The reason is that the long-term memory of LSTM contributes to recognizing the specific symbol sequence of the packet header, which means that it can use the correctly recognized symbol sequence in the header to assist in identifying the rest of the parts. Through increasing the header recognition accuracy, more packets can be collected, and the data rate goes higher.

Next, we investigate the sensitivity of BER and data rate performance to the identification ability of the encoder. Here, we build two CNN-based encoders for comparison, one is with the CAE-LSTM-Net-like structure but replacing the recalibration unit with a residual unit, and the other is with a structure comprised of basic convolutional layers, ReLU layers and BatchNormalizaiton layers such as LeNet-5, as shown in Figure 3. These encoders are named residual-Net and convolution-Net, respectively, and the scale of their weights are set to be the same level as that of channel-attention-Net for fair comparison. Figures 15 and 16 demonstrate the BER and data rate performance using encoders of different structures. As shown in Figure 15, the channel-attention-Net-based encoder significantly improves the BER performance in all cases of symbol rate. In a lower symbol rate, BER will not increase in a linear relationship with the increasing of symbol rate, and the residual-Net-based encoder can perform satisfactorily. However, when the symbol rate is higher than 40 kbaud/s, only the scheme with a channel-attention-Net-based encoder still meets the FEC requirement, which means that more sufficient feature extraction contributes to generating a descriptor with better quality and mitigating ISI. As shown in Figure 16, in the conditions of higher symbol rate, the channel-attention-Net based encoder outperforms in data rate. Better recognition of the header is the main reason for the increase of data rate.

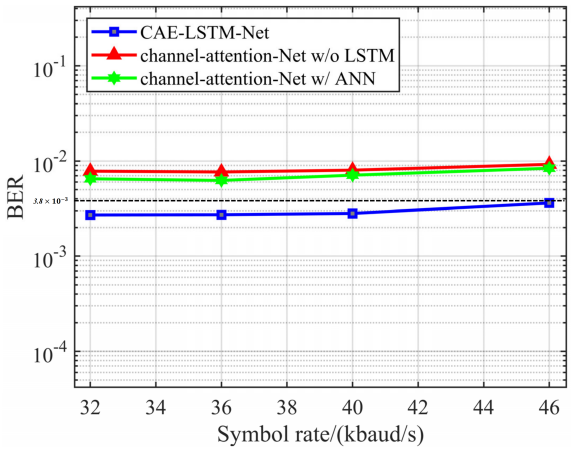

Figure 13. BER performance by using different equalizers.

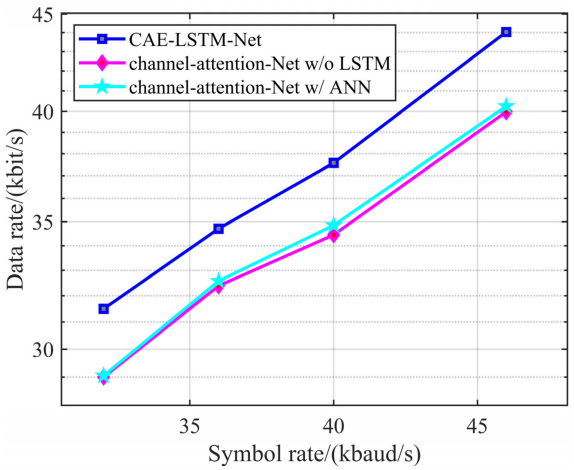

Figure 14. Data rate comparison by using different equalizers.

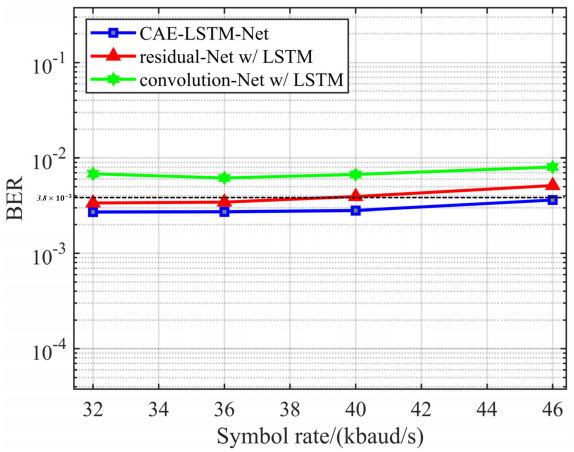

Figure 15. BER performance by using different CNN-based encoders.

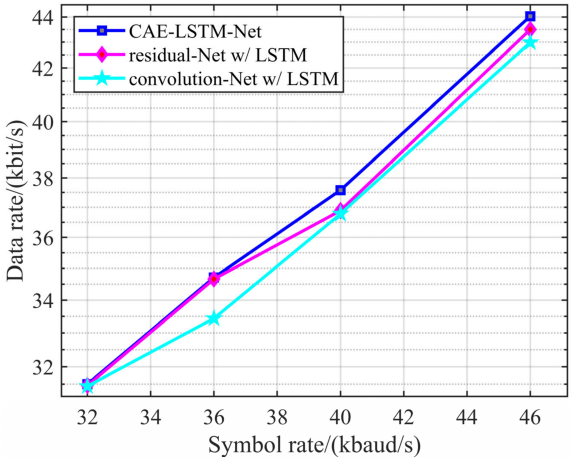

Figure 16. Data rate comparison by using different CNN-based encoders.

6. Conclusions

We have proposed a neural network structure based on channel-attention-Net and LSTM, which is named CAE-LSTM-Net for data decoding in an OCC system. Benefiting from the CNN's capability of extracting features, our proposed channel-attention-Net can generate a precise descriptor of input images. Transforming the decoding problem of a frame into a time-series classification problem, we proposed an LSTM-based equalizer to mitigate the performance deterioration caused by ISI. The experimental results demonstrate that our proposed CAE-LSTM-Net-based decoding scheme significantly improves the performance of data rate and BER in an OCC system. Based on the proposed decoding algorithms, a record-high data rate of 44.03 kbit/s is achieved by the RGB-LED-based OCC system, also meeting the FEC requirement.

Author Contributions: Conceptualization, W.G.; methodology, P.L.; software, P.L.; investigation, M.L.; writing—review and editing, P.L. and W.G.; supervision, W.G.; project administration, P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Younus, O.I.; Hassan, N.B.; Ghassemlooy, Z.; Haigh, P.A.; Zvanovec, S.; Alves, L.N.; Le Minh, H. Data rate enhancement in optical camera communications using an artificial neural network equaliser. *IEEE Access* **2020**, *8*, 42656–42665. [CrossRef]
2. Ashok, A. An Empirical Study of Deep Learning Models for LED Signal Demodulation in Optical Camera Communication. *Network* **2021**, *1*, 261–278.
3. Tran, H.Q.; Ha, C. Improved visible light-based indoor positioning system using machine learning classification and regression. *Appl. Sci.* **2019**, *9*, 1048. [CrossRef]
4. Cossu, G.; Sturniolo, A.; Ciaramella, E. Modelization and characterization of a CMOS camera as an optical real-time oscilloscope. *IEEE Photonics J.* **2020**, *12*, 1–13. [CrossRef]
5. Yu, K.; He, J.; Huang, Z. Decoding scheme based on CNN for mobile optical camera communication. *Appl. Opt.* **2020**, *59*, 7109–7113. [CrossRef] [PubMed]
6. *IEEE Standard 802.15.7-2011*; IEEE Standard for Local and Metropolitan Area Networks–Part 15.7: Short-Range Wireless Optical Communication Using Visible Light. IEEE: New York, NY, USA, 2011; pp. 1–309. [CrossRef]
7. Chen, H.W.; Wen, S.S.; Wang, X.L.; Liang, M.Z.; Li, M.Y.; Li, Q.C.; Liu, Y. Color-shift keying for optical camera communication using a rolling shutter mode. *IEEE Photonics J.* **2019**, *11*, 1–8. [CrossRef]
8. Chow, C.W.; Chen, C.Y.; Chen, S.H. Enhancement of signal performance in LED visible light communications using mobile phone camera. *IEEE Photonics J.* **2015**, *7*, 1–7. [CrossRef]
9. Chen, H.; Lai, X.; Chen, P.; Liu, Y.; Yu, M.; Liu, Z.; Zhu, Z. Quadrichromatic LED based mobile phone camera visible light communication. *Opt. Express* **2018**, *26*, 17132–17144. [CrossRef]
10. Hu, P.; Pathak, P.H.; Feng, X.; Fu, H.; Mohapatra, P. Colorbars: Increasing data rate of led-to-camera communication using color shift keying. In Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies, Heidelberg, Germany, 1–4 December 2015; pp. 1–13.

11. Li, J.; Guan, W. The optical barcode detection and recognition method based on visible light communication using machine learning. *Appl. Sci.* **2018**, *8*, 2425. [CrossRef]
12. Sun, X.; Shi, W.; Cheng, Q.; Liu, W.; Wang, Z.; Zhang, J. An LED detection and recognition method based on deep learning in vehicle optical camera communication. *IEEE Access* **2021**, *9*, 80897–80905. [CrossRef]
13. Dong, N.C.; Jin, S.Y.; Lee, J.; Kim, B.W. Deep Learning Technique for Improving Data Reception in Optical Camera Communication-Based V2I. In Proceedings of the 2019 28th International Conference on Computer Communication and Networks (ICCCN), Valencia, Spain, 29 July–1 August 2019.
14. Islam, A.; Hossan, M.T.; Jang, Y.M. Convolutional Neural Network Scheme-Based Optical Camera Communication System for Intelligent Internet of Vehicles. *Int. J. Distrib. Sens. Netw.* **2018**, *14*, 155014771877015. [CrossRef]
15. Liu, L.; Deng, R.; Chen, L.K. 47-kbit/s RGB-LED-based optical camera communication based on 2D-CNN and XOR-based data loss compensation. *Opt. Express* **2019**, *27*, 33840–33846. [CrossRef] [PubMed]
16. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
17. Jelodar, H.; Wang, Y.; Orji, R.; Huang, S. Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2733–2742. [CrossRef] [PubMed]
18. Li, S.; Yan, Z.; Wu, X.; Li, A.; Zhou, B. A method of emotional analysis of movie based on convolution neural network and bi-directional LSTM RNN. In Proceedings of the 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), Shenzhen, China, 26–29 June 2017; pp. 156–161.
19. Xu, Z.; Chen, T.; Qin, G.; Chi, N. Applications of Machine Learning in Visible Light Communication. In Proceedings of the 2021 18th China International Forum on Solid State Lighting & 2021 7th International Forum on Wide Bandgap Semiconductors (SSLChina: IFWS), Shenzhen, China, 6–8 December 2021; pp. 198–201.
20. Liu, Y. Decoding mobile-phone image sensor rolling shutter effect for visible light communications. *Opt. Eng.* **2016**, *55*, 016103. [CrossRef]
21. Chow, C.W.; Shiu, R.J.; Liu, Y.C.; Liu, Y.; Yeh, C.H. Non-flickering 100 m RGB visible light communication transmission based on a CMOS image sensor. *Opt. Express* **2018**, *26*, 7079–7084. [CrossRef]
22. Landis, C. Determinants of the critical flicker-fusion threshold. *Physiol. Rev.* **1954**, *34*, 259–286. [CrossRef]
23. Danakis, C.; Afgani, M.; Povey, G.; Underwood, I.; Haas, H. Using a CMOS camera sensor for visible light communication. In Proceedings of the 2012 IEEE Globecom Workshops, Anaheim, CA, USA, 3–7 December 2012; pp. 1244–1248.
24. Wang, W.C.; Chow, C.W.; Chen, C.W.; Hsieh, H.C.; Chen, Y.T. Beacon jointed packet reconstruction scheme for mobile-phone based visible light communications using rolling shutter. *IEEE Photonics J.* **2017**, *9*, 1–6. [CrossRef]
25. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
26. Hochreiter, S.; Bengio, Y.; Frasconi, P.; Schmidhuber, J. *Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies*; A Field Guide to Dynamical Recurrent Networks; Wiley-IEEE Press: Piscataway, NJ, USA, 2001; pp. 237–243.
27. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef]
28. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef] [PubMed]
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
30. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2874–2883.
31. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Las Vegas, NV, USA, 27–30 June 2016; pp. 483–499.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Santurkar, S.; Tsipras, D.; Ilyas, A.; Madry, A. How does batch normalization help optimization? *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 2488–2498.
34. Liu, Y.; Chow, C.W.; Liang, K.; Chen, H.Y.; Hsu, C.W.; Chen, C.Y.; Chen, S.H. Comparison of thresholding schemes for visible light communication using mobile-phone image sensor. *Opt. Express* **2016**, *24*, 1973–1978. [CrossRef] [PubMed]
35. Chow, C.W.; Chen, C.Y.; Chen, S.H. Visible light communication using mobile-phone camera with data rate higher than frame rate. *Opt. Express* **2015**, *23*, 26080–26085. [CrossRef]
36. Chow, C.W.; Liu, Y.; Yeh, C.H.; Chang, Y.H.; Lin, Y.S.; Hsu, K.L.; Liao, X.L.; Lin, K.H. Display Light Panel and Rolling Shutter Image Sensor Based Optical Camera Communication (OCC) Using Frame-Averaging Background Removal and Neural Network. *J. Light. Technol.* **2021**, *39*, 4360–4366. [CrossRef]
37. Liang, K.; Chow, C.W.; Liu, Y. RGB visible light communication using mobile-phone camera and multi-input multi-output. *Opt. Express* **2016**, *24*, 9383–9388. [CrossRef]

38. Chow, C.W.; Shiu, R.J.; Liu, Y.C.; Yeh, C.H.; Liao, X.L.; Lin, K.H.; Wang, Y.C.; Chen, Y.Y. Secure mobile-phone based visible light communications with different noise-ratio light-panel. *IEEE Photonics J.* **2018**, *10*, 1–6. [CrossRef]
39. Hsu, K.L.; Chow, C.W.; Liu, Y.; Wu, Y.C.; Hong, C.Y.; Liao, X.L.; Lin, K.H.; Chen, Y.Y. Rolling-shutter-effect camera-based visible light communication using RGB channel separation and an artificial neural network. *Opt. Express* **2020**, *28*, 39956–39962. [CrossRef]
40. Aoyama, H.; Oshima, M. Visible light communication using a conventional image sensor. In Proceedings of the 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, 9–12 January 2015; pp. 103–108.