

Temporal-context 1D convolutional neural network for joint modulation and transmission-distance classification in visible light communication

Boyong Liu¹, Qiyue Zhang^{2*}

¹School of Electronic and Information Engineering, Beijing Jiaotong University (BJTU), Beijing 100044, China

²Navigation Aids Maintenance Division, North China Regional Air Traffic Management Bureau, Beijing 100621, China

Article info

Article history:

Received 26 Aug. 2026

Received in revised form 08 Sep. 2026

Accepted 14 Sep. 2026

Available on-line 08 Oct. 2026

Keywords:

visible light communication;

deep learning;

1D-CNN;

modulation classification;

distance classification.

Abstract

Accurate identification of modulation format and transmission distance is important for reliable visible light communication (VLC) systems, particularly under channel-dependent signal distortions. This study presents a temporal-context one-dimensional convolutional neural network (1D-CNN) for simultaneous modulation-format recognition and transmission-distance classification using experimentally acquired VLC signals. The proposed network employs four convolutional layers with batch normalisation and Leaky ReLU activation, followed by global average pooling, a fully connected layer, and two softmax output heads for modulation-format and transmission-distance classification. A signal segment length of $N = 40$ and a temporal-context of $K = 10$ were selected based on comparative experiments. The framework was evaluated for seven modulation formats and 29 discrete transmission-distance classes from 0 to 140 cm at 5-cm intervals, achieving 99.5% modulation-format classification accuracy and 98.7% distance-classification accuracy. The distance results demonstrate discrete identification of experimentally measured link positions rather than general-purpose distance ranging. The results further show that classification performance decreases for higher-order modulation formats and longer transmission distances because of increased constellation density and channel distortion. The proposed approach provides a compact solution for joint signal-format and distance identification directly from received temporal waveforms.

1. Introduction

Visible light communication (VLC) is a form of optical wireless communication (OWC) technology that uses visible light to transmit information wirelessly. VLC can exploit existing LED lighting infrastructure, providing a communication mechanism that does not require a dedicated physical transmission medium such as optical fibre. Unlike fibre-optic communication, VLC provides a wireless optical link in which information is transmitted through free space rather than through a guided optical medium [1–5]. In addition, the spatial confinement of visible light offers inherent advantages in communication

security and interference isolation. These characteristics have made VLC an attractive technology for short-range wireless communication and indoor networking applications [6–9]. VLC has evolved from early optical wireless communication concepts to modern LED-based communication systems [10–13].

At the physical layer, channel attenuation, multipath propagation, noise, and other signal impairments can affect reliable VLC reception. Conventional communication receivers are commonly designed under assumed channel and noise models and may require channel-state information (CSI) and knowledge of channel characteristics for reliable detection [14–16]. However, practical wireless channels can exhibit multipath fading, impulse noise, interference, and other non-ideal disturbances

*Corresponding author at: qiyue_zhang@bjtu.edu.cn

that can significantly degrade signal detection and demodulation performance [17]. These limitations motivate the investigation of data-driven approaches that learn discriminative signal features directly from received waveforms. Recent studies have demonstrated the potential of deep learning for end-to-end VLC transmission and for mitigating practical channel and transceiver impairments [18, 19]. Therefore, researchers have investigated machine learning (ML) and deep learning (DL) as data-driven alternatives to conventional model-dependent signal processing. Previous studies have used neural networks for wireless-channel estimation and automatic modulation detection [20, 21]. These approaches indicate that learned representations can capture signal and channel characteristics without requiring an explicit analytical model of the complete communication channel. In practical VLC systems, the receiver may not have access to the modulation format through explicit signalling information. This situation can arise in blind reception, adaptive or reconfigurable VLC links, cognitive receivers, loss of control or signalling information, or scenarios involving multiple transmitters with unknown configurations. In such cases, automatic modulation-format recognition can provide the receiver with information about the modulation scheme directly from the received waveform, enabling subsequent signal processing and demodulation without relying on prior knowledge of the format. Therefore, this work considers modulation-format recognition defined as a receiver-side capability for signal identification when format information is unknown or unavailable.

In VLC, researchers have explored ML and DL techniques for several signal-processing tasks. Recent surveys have shown that learning-based methods have been investigated for channel estimation, non-linear compensation, modulation detection, positioning, and other VLC physical-layer tasks [22]. CNN-based approaches have been investigated for reducing linear and non-linear distortion in VLC non-orthogonal multiple access (VLC-NOMA) systems [23–25]. Clustering-based methods have been reported for classification and non-linear compensation in VLC systems employing pulse-amplitude modulation [26, 27]. Recurrent neural networks, including long short-term memory (LSTM) architectures, have also been investigated to compensate for transmission disturbances in PAM-based VLC links [28]. In addition, deep-learning-based approaches have been applied to multi-colour VLC systems [29], while combined learning-based channel estimation and decision strategies have also been investigated [30]. Despite these developments, several aspects of learning-based VLC signal processing remain comparatively underexplored. Recent studies have demonstrated deep-learning-based modulation recognition and, more recently, joint modulation classification and distance estimation in VLC systems [31, 32].

To address these issues, this study investigates a one-dimensional CNN-based framework that operates directly on received VLC signal samples for modulation and transmission-distance classification. Rather than converting the received signals into image representations, the proposed approach uses the signal itself as the network input. In addition, we jointly incorporate consecutive signal segments to investigate whether temporal context can improve classification performance. Six temporal-context

configurations, namely $K=1, 3, 5, 8, 10$, and 12 , are evaluated. To address these issues, this study investigates a one-dimensional convolutional neural network (1D-CNN)-based framework that operates directly on received VLC signal samples for modulation and transmission-distance classification. Rather than converting the received signals into image representations, the proposed approach uses the signal itself as the network input. In addition, consecutive signal segments are jointly incorporated to investigate whether temporal context can improve classification performance. Six temporal-context configurations, namely $K=1, 3, 5, 8, 10, 12$, are evaluated. The $K=10$ configuration is subsequently selected based on the classification performance observed among the investigated configurations and is used for the analysis of different modulation formats and transmission distances. The main contributions of this study are as follows: (1) A 1D-CNN framework is developed for classification of modulation format and transmission distance directly from received VLC signal samples, avoiding image-based signal representation. (2) A temporal-context input strategy is investigated by jointly incorporating consecutive signal segments into the CNN input, with six configurations of $K=1, 3, 5, 8, 10, 12$. (3) The effect of temporal context on classification performance is systematically evaluated, and $K=10$ is selected based on the experimentally observed performance among the tested configurations. (4) The proposed approach is evaluated for seven modulation formats – on-off keying (OOK), quadrature phase shift keying (QPSK), 16-QAM, 32-QAM, 64-QAM, 128-QAM, and 256-QAM – over transmission distances ranging from 0 to 140 cm. (5) The dependence of classification performance on modulation order and transmission distance is investigated to characterise the operating behaviour of the proposed CNN-based approach.

2. Proposed algorithm

This section presents the proposed deep-learning framework for joint modulation-format and transmission-distance classification in a VLC system. The methodology includes four main stages: describing the experimental VLC system and dataset, signal pre-processing and temporal-context construction, developing the 1D-CNN, and training and joint classification. The complete processing framework operates directly on experimentally acquired received waveforms without converting the signals into images or other time-frequency representations.

2.1. Experimental VLC system and dataset

The VLC system considered in this study consists of a transmitter, a free-space optical channel, a receiver, and a machine-learning-based signal-processing stage. Figure 1 illustrates the overall experimental configuration. At the transmitter, the original digital signal $[S(n)]$ is first modulated using the selected modulation format. The resulting waveform is generated as an analogue electrical signal using an arbitrary waveform generator (AWG). The electrical signal is subsequently amplified and combined with a DC bias through a bias-T circuit to provide the appropriate operating condition for the LED. The biased electrical signal drives the LED, which converts

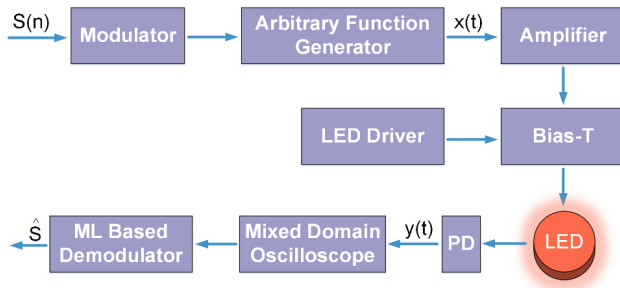


Fig. 1. Experimental VLC system and signal-processing chain.

the electrical waveform into an intensity-modulated optical signal. The optical signal propagates through the free-space VLC channel and is detected by a photodiode (PD) at the receiver. The PD converts the received optical signal into an electrical signal, which is then digitised using a mixed-domain oscilloscope. The resulting digital waveform is then provided to the machine-learning-based processing stage for modulation and transmission-distance classification.

The dataset employed in this work is a publicly available dataset generated from a flexible hardware-based VLC prototype. The dataset contains received signals experimentally acquired from the VLC system and subsequently prepared and normalised for machine-learning applications. Therefore, the dataset represents experimentally measured VLC waveforms rather than synthetically generated signals. The dataset is publicly available through the source reported in [23]. Normalisation procedure. The received signals used in this study follow the normalisation procedure described for the original dataset in [23]. Before machine-learning processing, the received sampled data are normalised to the range $[0,1]$ using the minimum and maximum values of the corresponding received-data set. Specifically, for a received sample $\hat{y}_i$, the normalised value is defined as (1):

$$\hat{y}_i = (\hat{y}_i - \hat{y}_{\min}) / (\hat{y}_{\max} - \hat{y}_{\min}), \quad (1)$$

where $\hat{y}_{\min}$ and $\hat{y}_{\max}$ denote the minimum and maximum values used in the normalisation procedure. Thus, the normalisation is not performed independently for each individual waveform or temporal-context window. This distinction is important because independent per-waveform normalisation would remove the absolute amplitude scale of each waveform, whereas the normalisation procedure adopted in the source dataset retains relative amplitude differences among the received signals. Consequently, distance-dependent amplitude variations may remain available to the classifier as a discriminative cue. The normalisation parameters are computed using the entire available dataset before partitioning into training, validation, and test subsets, consistent with the pre-processing described in [23].

The original dataset contains several modulation formats, including OOK, QPSK, 16-QAM, 32-QAM, 64-QAM, 128-QAM, 256-QAM, and 4-PPM. In the present study, seven modulation formats, namely OOK, QPSK, 16-QAM, 32-QAM, 64-QAM, 128-QAM, and 256-QAM, are investigated. The signals were experimentally collected for transmission distances ranging from 0 to 140 cm with a spatial interval of 5 cm. The dataset includes

four signal-segment configurations, corresponding to $(N=10)$, 20, 40, and 80 samples. Here, (N) denotes the number of sampled points within one signal period, rather than an arbitrary number of samples per signal period (symbol interval). According to [23], the experimental arbitrary waveform generator operated at a sampling rate of 2.5 MS/s. Therefore, for the $(N=40)$ configuration used in the present study, each signal segment contains 40 consecutive samples corresponding to a signal period of $16 \mu\text{s}$. If this signal period is taken as the symbol interval, the corresponding symbol rate is 62.5 kSymbols/s. Each N -sample signal segment is associated with a corresponding symbol-state label in the original dataset. Thus, the received waveform is first represented as consecutive N -sample signal segments, after which multiple consecutive segments are combined to construct the temporal-context input. Each signal segment is associated with a corresponding symbol-state label within the known modulation format. For example, 16-QAM contains 16 symbol states, represented by labels from 0 to 15, whereas 32-QAM contains 32 symbol states, represented by labels from 0 to 31. These labels identify the transmitted symbol states and therefore serve as the target classes for symbol demodulation when the modulation format is fixed. They should not be interpreted as modulation-format labels. For example, an experiment performed exclusively with 32-QAM and labels 0–31 represents 32-class symbol demodulation, rather than seven-class modulation-format recognition. Figures 2 and 3 show representative examples of the received waveforms. For QPSK, the transmitted symbols are generated from two orthogonal carrier components: the in-phase and quadrature components. These components can be represented by sinusoidal and cosine waveforms with a phase difference of (90°) . Different QPSK symbols are therefore distinguished primarily by their phase relationships. Consequently, the waveform associated with each symbol label exhibits a different phase state while maintaining the corresponding carrier structure.

For 16-QAM, the transmitted symbols are determined jointly by the amplitude and phase of the carrier components. As illustrated in Fig. 3, although 16 symbol states are available, only representative symbol labels are shown for clarity. The in-phase and quadrature components remain orthogonal, with a (90°) phase difference,

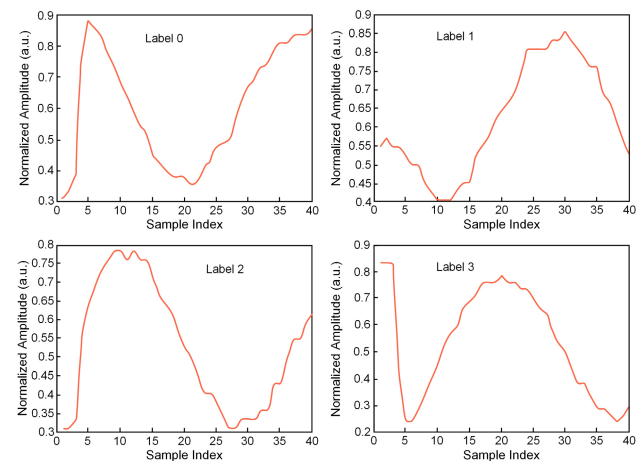


Fig. 2. Representative received QPSK signal waveforms corresponding to different symbol labels.

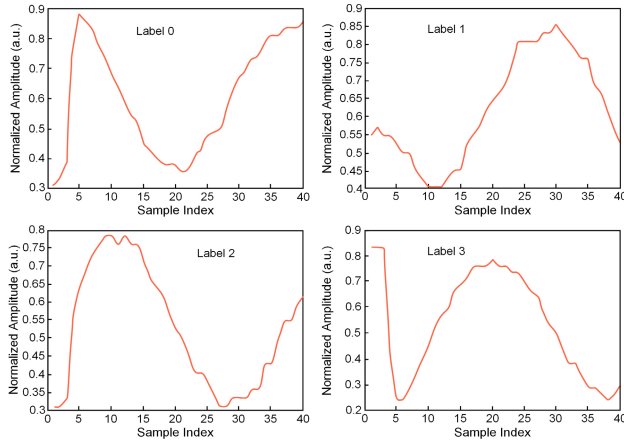


Fig. 3. Representative received 16-QAM signal waveforms for selected symbol labels.

while their amplitudes vary according to the corresponding symbol. Consequently, the received 16-QAM waveforms exhibit variations in both amplitude and phase.

The same signal-processing procedure is applied to the remaining investigated modulation formats. Since the four available values of (N) produce different input representations and classification performances, the appropriate symbol interval is first investigated experimentally. The preliminary evaluation uses 32-QAM as the representative modulation format. The results of this investigation are used to select the symbol interval employed in the subsequent experiments. For the selected ($N=40$) configuration, Table 1 summarises the signal matrix dimensions and corresponding symbol-label ranges for the seven investigated modulation formats.

Table 1.
Signal dimensions and symbol-label ranges for the seven investigated modulation formats.

Modulated signal	Signal length	Labels
16-QAM	40×67500	0–15
32-QAM	40×36000	0–31
64-QAM	40×72000	0–63
128-QAM	40×72000	0–127
256-QAM	40×72000	0–255
OKK	40×18000	0–1
QPSK	40×18000	0–3

Note: The labels denote symbol states within each known modulation format and are used as targets in the symbol-demodulation experiments; they do not represent modulation-format classes.

The first dimension of each matrix denotes the number of samples in one signal segment ($N=40$), while the second dimension denotes the number of available signal periods (signal segments) in the corresponding dataset. Thus, for example, $40 \times 36\,000$ represents 36 000 signal segments, each containing 40 samples. These experimentally acquired segments are subsequently used to construct the temporal-context inputs described in section 2.2.

2.2. Signal preprocessing and temporal-context construction

The received waveforms are first segmented into fixed-length signal units according to the selected symbol interval, denoted by (N). Preliminary experiments are initially conducted using 32-QAM to determine the appropriate symbol interval. This modulation format is selected as a representative case for evaluating the effect of symbol interval on classification performance before extending the analysis to the other investigated modulation formats. For each candidate value of (N), the received waveform is divided into consecutive signal segments, and the resulting segments are supplied to a 1D-CNN for classification. Four symbol intervals, ($N=10$), (20), (40), and (80), are considered. The classification performance obtained for these configurations is compared, and the symbol interval providing the most suitable classification performance is selected for the subsequent experiments. Based on this preliminary evaluation, ($N=40$) is selected and subsequently used throughout the main experiments.

After determining the symbol interval, temporal information between consecutive signal segments is incorporated into the input representation. After determining the symbol interval, temporal information from consecutive signal segments is incorporated into the input representation. This choice is motivated by the fact that practical VLC channels can contain direct, reflected, and scattered optical components with different propagation paths. Such channel characteristics can produce waveform distortion and temporal spreading in the received signal. Consequently, neighbouring received signal segments may contain related information useful for classification. However, the present experiments do not isolate multipath propagation from other factors that may contribute to the observed classification improvement. Therefore, the temporal-context analysis is interpreted as an investigation of whether jointly processing consecutive signal segments improves classification performance, rather than as a direct experimental proof that the improvement is caused specifically by multipath-induced temporal memory.

To exploit temporal dependencies between consecutive signal segments, the network jointly receives multiple (N)-sample segments. The temporal context is denoted by (K), where ($K \in \{1, 3, 5, 8, 10, 12\}$), and represents the number of consecutive signal segments included in one input sample. For example, the ($K=10$) configuration jointly incorporates ten consecutive signal segments into a single temporal-context input. More specifically, let (X_i) denote the i -th signal segment, where each (X_i) is a vector of length ($N=40$) samples (one symbol interval). The temporal-context input associated with the i -th segment is constructed as (2):

$$X_i^{(K)} = [x_{i-[K/2]}, x_i, \dots, x_{i+[K/2]}], \quad (2)$$

where (K) represents the number of consecutive signal segments incorporated into the input and the index range is truncated at the segment boundaries when necessary. With ($N=40$), the selected ($K=10$) configuration contains ten consecutive segments, each consisting of 40 samples. The input is arranged as a (40×10) matrix, where the first dimension represents the samples within each signal

segment and the second dimension represents the ($K = 10$) consecutive segments. In other words, the input is formed as 10 separate channels, each containing 40 samples from a different symbol interval, not as a single sequence of 400 consecutive samples. The 1D convolutional layers operate along the first dimension (40 samples), while the second dimension (10 channels) provides multi-channel temporal context. Thus, ($K = 10$) represents a 10-segment temporal context rather than a convolutional dimensionality. To avoid ambiguity, the configurations are referred to throughout the manuscript as ($K = 1$), ($K = 3$), ($K = 5$), ($K = 8$), ($K = 10$), and ($K = 12$), rather than $K = 1, 3, 5, 8, 10, 12$. The complete pre-processing and temporal-context construction procedure is illustrated in Fig. 4.

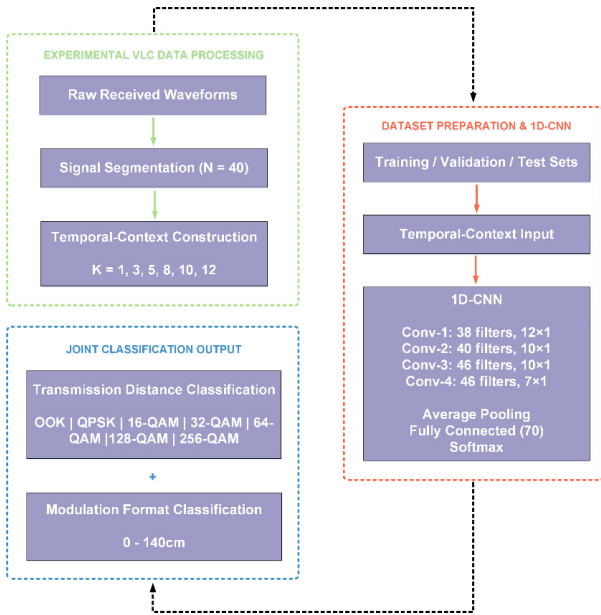


Fig. 4. Signal pre-processing and temporal-context construction for the proposed modulation- and distance-classification framework.

The six temporal-context configurations are subsequently evaluated using the same CNN framework. This allows the contribution of temporal information to modulation and transmission-distance classification to be examined systematically. The $K = 10$ configuration is selected based on the classification performance observed among the tested temporal-context configurations, as discussed in section 3.2.

2.3. Proposed 1D-CNN architecture

A 1D-CNN extracts discriminative features directly from the received VLC waveforms. Unlike image-based deep-learning approaches, the proposed method does not require converting the received signals into spectrograms, constellation images, or other 2D representations. The network therefore operates directly on the temporal structure of the experimentally acquired signal. The CNN architecture is initially developed using 32-QAM data. Several combinations of convolutional filter sizes and numbers of filters are investigated to determine an appropriate network configuration. Based on these preliminary experiments, a four-layer 1D convolutional architecture is adopted. The proposed architecture of four-layer 1D-CNN is shown in Fig. 5. The input consists of temporal-context signal segments with a segment length of $N = 40$ and a temporal-context of $K = 10$. The network comprises four 1D convolutional layers with kernel sizes of 12×1 , 10×1 , 10×1 , and 7×1 , and 38, 40, 46, and 46 filters, respectively. Each convolutional layer is followed by batch normalisation and Leaky ReLU activation ($\alpha = 0.01$). The extracted feature maps are aggregated using global average pooling and subsequently processed by a fully connected layer with 70 units and a softmax classifier. The network has two separate output heads: (i) a modulation-format classification head that predicts one of seven modulation formats (OOK, QPSK, 16-QAM, 32-QAM, 64-QAM, 128-QAM, or 256-QAM), and (ii) a transmission-distance classification head that predicts one of 29 distance classes (0–140 cm with a 5 cm interval).

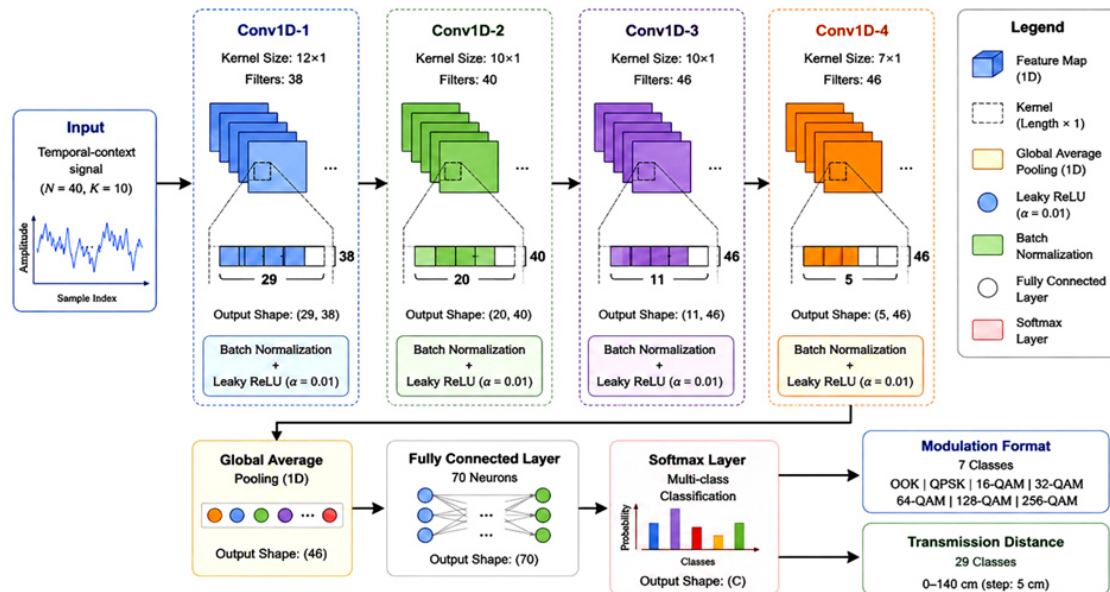


Fig. 5. Architecture of the proposed four-layer 1D-CNN for joint modulation-format and transmission-distance classification.

Note that the network does not include a symbol-demodulation output in this final joint configuration. The symbol-demodulation experiments described in sections 3.1–3.2 are performed separately during the architecture-optimisation phase, using a network with a single output head for the 32 symbol classes of 32-QAM. All convolutional layers use a stride of 1 and same padding.

The proposed network consists of four successive convolutional layers. The first convolutional layer contains 38 filters with a filter size of (12×1) , while the second layer contains 40 filters with a filter size of (10×1) . The third convolutional layer contains 46 filters with a filter size of (10×1) , and the fourth layer contains 46 filters with a filter size of (7×1) . The selected filter sizes and numbers of filters were determined through preliminary network-development experiments. The CNN architecture parameters are summarised in Table 2.

Table 2.

CNN architecture parameters used in the proposed framework.

Layers	Filter size	Number of filters
Conv-1	12×12	38
Conv-2	10×10	40
Conv-3	10×10	46
Conv-4	7×7	46

The convolutional layers progressively transform the raw temporal waveform into higher-level feature representations. Early convolutional layers extract local waveform characteristics, whereas deeper layers combine these local features to obtain more discriminative representations for modulation and distance classification. After the four convolutional layers, an average-pooling operation aggregates the extracted feature maps and reduces their effective dimensionality. Average pooling calculates the mean value of the features covered by the corresponding pooling region. This operation reduces the computational requirements of the subsequent classification

layers while retaining the dominant information extracted by the convolutional stages. The pooled representation is then fed into a fully connected layer with 70 units. This layer performs the final feature mapping before classification. The output layer employs the softmax activation function to transform the network outputs into class probabilities. The probability corresponding to the maximum output is selected as the predicted class. The complete network architecture and its relationship with the signal-processing and classification stages are further illustrated in Fig. 6. The framework consists of three main stages: (1) signal pre-processing, including signal segmentation with a segment length of $N=40$, temporal-context construction with $K \in \{1, 3, 5, 8, 10, 12\}$, and dataset partitioning into training (70%), validation (15%), and test (15%) sets; (2) the proposed 1D-CNN feature-extraction network, comprising four 1D convolutional layers followed by batch normalisation and Leaky ReLU activation ($\alpha=0.01$), global average pooling, and a fully connected layer with 70 units; and (3) two classification heads for joint prediction of the modulation format and transmission distance. The modulation head predicts seven modulation classes, whereas the distance head predicts 29 transmission-distance classes ranging from 0 to 140 cm with a 5-cm interval. All convolutional layers employ a stride of 1 and same padding.

The network architecture shown in Fig. 6 therefore provides the main feature-extraction and classification framework used throughout this study. The same basic CNN structure is used for the different temporal-context configurations so the effect of temporal information can be evaluated without changing the underlying network architecture. The shared feature representation, after global average pooling and the 70-unit fully connected layer, is mapped to two task-specific output heads. The modulation-format head consists of a fully connected layer with dimensions $[70 \times 7]$ followed by a softmax activation, producing probabilities over the seven modulation classes (OOK, QPSK, 16-QAM, 32-QAM, 64-QAM, 128-QAM, and 256-QAM). The transmission-distance head consists of a fully connected layer with dimensions $[70 \times 29]$ followed

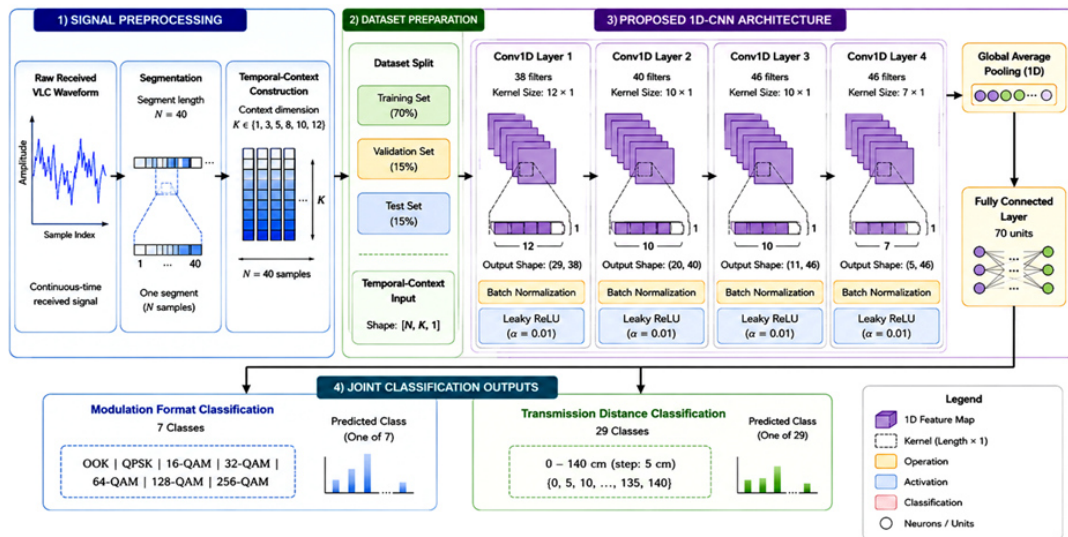


Fig. 6. Overview of the proposed temporal-context 1D-CNN framework, including signal segmentation into ($N=40$) samples per signal period, block-wise dataset partitioning into training (70%), validation (15%), and test (15%) subsets, temporal-context construction with ($K \in \{1, 3, 5, 8, 10, 12\}$) performed independently within each subset, and subsequent joint modulation-format and transmission-distance classification.

by a softmax activation, producing probabilities over the 29 distance classes (0–140 cm, 5-cm intervals). Both heads share the same feature extractor, enabling the network to learn representations that benefit both tasks simultaneously.

2.4. Training and joint modulation–distance classification procedure

The proposed CNN is trained using the Adam optimisation algorithm. A mini-batch size of 3500 is employed, and the maximum number of training epochs is selected within the range of 100–200 depending on the experimental configuration. The corresponding training parameters are summarised in Table 3.

Table 3.
Training parameters of the proposed CNN.

Parameter	Value
Maximum number of epochs	100–200
Mini-batch size	3500
Optimiser	Adam

Data partitioning and prevention of data leakage. To prevent data leakage arising from overlapping temporal-context windows, dataset partitioning is performed before temporal-context construction. The underlying N -sample signal segments are first divided into non-overlapping chronological blocks and assigned to the training (70%), validation (15%), and test (15%) subsets. Temporal-context windows are subsequently generated independently within each subset using a stride of one signal segment. Thus, although consecutive windows within the same subset may share $(K-1)$ signal segments, no temporal-context window crosses a subset boundary, and no underlying signal segment is shared between different subsets. The validation subset is used for model and hyperparameter selection, including the number of samples per signal period (N), temporal context (K), and CNN architecture parameters. The test subset is reserved

exclusively for the final performance evaluation and is not used during model development or hyperparameter selection. This block-wise partitioning procedure is applied consistently to both the preliminary 32-QAM symbol-demodulation experiments and the final joint modulation-format and transmission-distance classification experiments.

To clarify the distinction between the different experimental tasks, Table 4 summarises the input data, target classes, network outputs, and evaluation metrics used throughout the study. In particular, the preliminary experiments based on 32-QAM evaluate symbol demodulation using the 32 transmitted symbol states as target classes, whereas the final experiments evaluate modulation-format recognition and transmission-distance classification using seven modulation-format classes and 29 distance classes, respectively.

The joint network is trained using two categorical cross-entropy objectives corresponding to the modulation-format and transmission-distance classification tasks. The modulation-format loss is defined as (3):

$$\mathcal{L}_{\text{mod}} = -\sum_{c=1}^7 y_c^{\text{mod}} \log(\hat{y}_c^{\text{mod}}), \quad (3)$$

where y_c^{mod} and $\hat{y}_c^{\text{mod}}$ denote the ground-truth and predicted probabilities for the c -th modulation-format class, respectively. Similarly, the transmission-distance loss is defined as (4):

$$\mathcal{L}_{\text{dist}} = -\sum_{d=1}^{29} y_d^{\text{dist}} \log(\hat{y}_d^{\text{dist}}), \quad (4)$$

where y_d^{dist} and $\hat{y}_d^{\text{dist}}$ denote the ground-truth and predicted probabilities for the d -th transmission-distance class, respectively.

The overall training objective is expressed as (5):

$$\mathcal{L}_{\text{total}} = \lambda_{\text{mod}} \mathcal{L}_{\text{mod}} + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}}, \quad (5)$$

where λ_{mod} and λ_{dist} are weighting coefficients that determine the relative contribution of the modulation-

Table 4.
Summary of experimental tasks, target classes, network outputs, and evaluation metrics.

Experiment	Input data	Target classes	Task	Output / evaluation	Metric
Value optimisation	32-QAM	32 symbol states (0–31)	Symbol demodulation	Symbol class	Symbol-demodulation accuracy
CNN architecture optimisation	32-QAM	32 symbol states (0–31)	Symbol demodulation	Symbol class	Symbol-demodulation accuracy
Temporal-context optimisation	32-QAM	32 symbol states (0–31)	Symbol demodulation	Symbol class	Symbol-demodulation accuracy
AdaBoost comparison	32-QAM	32 symbol states (0–31)	Symbol demodulation	Symbol class	Symbol-demodulation accuracy
Seven-format evaluation	OOK, QPSK, 16-QAM, 32-QAM, 64-QAM, 128-QAM, 256-QAM	Modulation format	Modulation-format recognition	Modulation format	Modulation-format classification accuracy
Distance evaluation	Seven modulation formats, 0–140 cm	29 distance classes	Transmission-distance classification	Distance class	Distance-classification accuracy

Note: The AdaBoost comparison refers to the 32-QAM symbol-demodulation results reported in [23]. It is not a comparison of seven-class modulation-format recognition. Because the two studies use different model architectures and input constructions, this comparison serves as a reference to a reported symbol-demodulation performance rather than a controlled one-to-one benchmark

format and transmission-distance losses, respectively. The training procedure has two main stages. First, the network-development experiments are performed using 32-QAM to determine the appropriate symbol interval and network configuration. The candidate symbol interval ($N=10$), 20, 40, and 80 are evaluated, and ($N=40$) is selected for the main experiments based on the resulting classification performance. The four-layer CNN architecture is then established by evaluating different filter configurations. In the second stage, the selected network is applied to the received signals collected at transmission distances from 0 to 140 cm. Initially, each distance is considered separately to evaluate the network modulation-classification capability under a fixed propagation condition. The classification performance is then evaluated across all transmission distances to determine the effect of propagation distance on the received-signal characteristics and classification accuracy. The initial single-context configuration processes each received signal segment independently. However, because VLC propagation may introduce delayed components through multipath transmission, the current signal segment can contain information related to preceding signal intervals. Therefore, the temporal-context configurations of $K=3, 5, 8, 10, 12$ are investigated in addition to the baseline $K=1$ configuration. The initial single-context configuration processes each received signal segment independently. To examine whether additional neighbouring waveform information is beneficial, temporal-context configurations of $K=3, 5, 8, 10, 12$ is investigated in addition to the baseline $K=1$ configuration. The purpose of increasing K is to provide the CNN with multiple consecutive received segments rather than only a single segment. This design allows the classifier to exploit information distributed across neighbouring segments. The observed performance improvement may reflect the additional waveform information available from the longer input context and may also be influenced by temporal correlations introduced by the propagation channel. Because the present study does not include a control experiment using shuffled or non-consecutive segments with the same number of input samples, the results do not uniquely establish multipath propagation as the cause of the improvement. Accordingly, the temporal-context analysis is interpreted as experimental evidence that jointly processing multiple consecutive segments can improve classification performance for the investigated dataset. The $K=10$ configuration is selected as the best-performing configuration among those tested in this study. The evaluation procedure distinguishes between symbol demodulation and modulation-format recognition. In the preliminary experiments using only 32-QAM data, the target classes are the 32 transmitted symbol states (0–31), so the resulting metric is symbol-demodulation accuracy. Since the modulation format is fixed in these experiments, these results do not represent modulation-format recognition. For the experiments involving the complete set of seven modulation formats, the modulation-format classification performance is evaluated according to the modulation identity associated with each test sample, while transmission-distance performance is evaluated according to the corresponding distance class. The resulting metrics are therefore reported separately as modulation-format

classification accuracy and transmission-distance classification accuracy. For modulation-format classification, the predicted modulation class from the modulation-format output head is compared with the ground-truth modulation identity of the corresponding test sample. The modulation-format classification accuracy is calculated as (6):

$$A_{\text{mod}} = \frac{N_{\text{correct, mod}}}{N_{\text{total}}} \times 100\%, \quad (6)$$

where $N_{\text{correct, mod}}$ is the number of correctly classified modulation-format samples and N_{total} is the total number of evaluated samples. The transmission-distance classification accuracy is similarly calculated by comparing the predicted distance class from the distance output head with the corresponding ground-truth distance class, as given in (7):

$$A_{\text{dist}} = \frac{N_{\text{correct, dist}}}{N_{\text{total}}} \times 100\%. \quad (7)$$

For each test interval, the predicted modulation-format and transmission-distance classes are compared with their corresponding ground-truth classes, and the number of correctly classified samples is determined. The resulting modulation-format and transmission-distance accuracies are calculated for individual transmission distances and for the complete test dataset. The overall methodology therefore combines experimentally acquired VLC waveforms, fixed-length temporal segmentation, temporal-context construction, four-layer 1D convolutional feature extraction, average pooling, fully connected classification, and softmax-based decision making.

3. Results and discussion

The performance of the proposed framework is experimentally evaluated from four complementary perspectives: symbol interval selection, CNN architecture optimisation, temporal-context analysis, joint modulation-format, and transmission-distance classification. The experiments are conducted using the experimentally acquired VLC dataset described in section 2. The preliminary 32-QAM results are first used to determine the appropriate input length and network configuration. The resulting model is then evaluated using the seven investigated modulation formats and the complete range of transmission distances.

3.1. Optimisation of symbol interval and CNN architecture

The first set of experiments investigates the symbol-demodulation performance of the proposed 1D-CNN using 32-QAM signals at different transmission distances. Because all samples in this stage use the same 32-QAM modulation format, the modulation format is known *a priori* and is not classified. Instead, the network predicts one of the 32 transmitted symbol states (labels 0–31). Therefore, the accuracies reported in this subsection represent 32-QAM symbol-demodulation accuracy, rather than modulation-format recognition accuracy. The initial network is trained separately for the investigated distances using the 1D signal representation. For short transmission distances, particularly from 0 to 40 cm, the network

achieves classification accuracy approaching 100%. However, performance degrades markedly as the transmission distance increases. This behaviour is consistent with the increasing influence of channel attenuation, waveform distortion, and multipath propagation on the received signal. The initial performance is presented in Fig. 7. The reduction in classification accuracy at larger distances indicates that a CNN optimised for short-range signals does not necessarily provide sufficient feature discrimination under more severe channel conditions. Therefore, the convolutional filter configuration is systematically investigated to improve the representation of distance-dependent waveform characteristics.

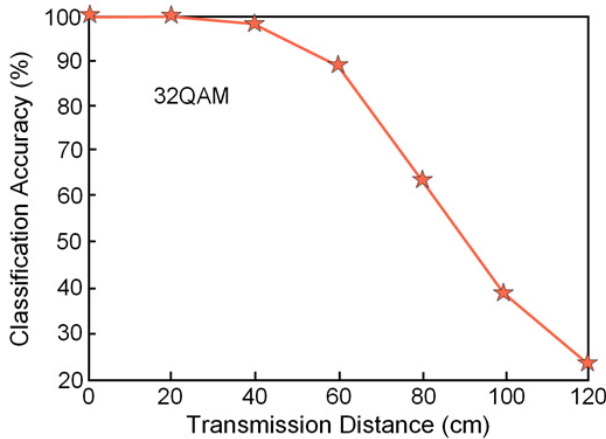


Fig. 7. Symbol-demodulation accuracy of the initial 1D-CNN for 32-QAM signals at different transmission distances. Because the modulation format is fixed to 32-QAM, the classification task corresponds to prediction of the 32 transmitted symbol states rather than modulation-format recognition.

Several combinations of filter sizes and numbers of convolutional filters are evaluated. One of the initial configurations is summarised in Table 5. The network contains four convolutional layers with progressively increasing numbers of filters, followed by a dense layer. Although this configuration improves feature extraction compared with the initial network, performance at longer transmission distances remains limited.

Table 5.

Initial CNN configuration evaluated for 1D 32-QAM symbol demodulation.

Layers	Filter size	Number of filters
Conv-1	3×3	30
Conv-2	5×5	30
Conv-3	5×5	64
Conv-4	5×5	64
Dense = 64	Epochs = 150	–

Further architectural optimisation is subsequently performed by varying both the receptive-field size and the number of filters. Table 6 presents the optimised configuration. Compared with the initial architecture, the optimised network employs larger convolutional receptive fields in the first layers and a reduced number of filters in the deeper layers. This configuration provides a more

effective representation of the received waveform temporal structure.

Table 6.

Optimised CNN configuration for 1D 32-QAM symbol demodulation.

Layers	Filter size	Number of filters
Conv-1	12×12	30
Conv-2	10×10	30
Conv-3	10×10	64
Conv-4	7×7	64
Dense = 70	Epochs = 150	–

The resulting classification performance is shown in Fig. 8. The optimised network provides a substantial improvement over the initial configuration, particularly for the more challenging transmission-distance conditions. This result demonstrates that the receptive field of the convolutional layers has an important influence on the ability of the network to extract discriminative features from distorted VLC waveforms.

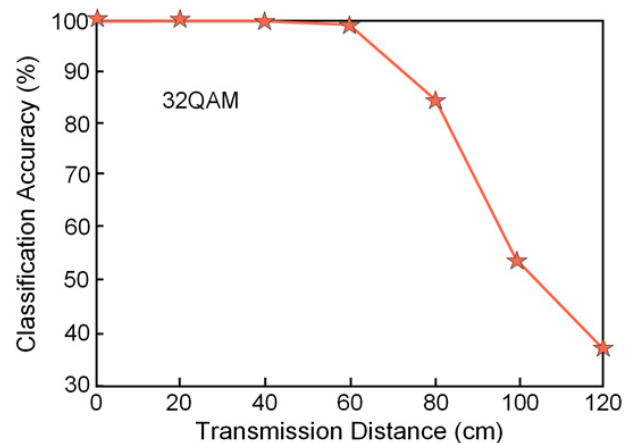


Fig. 8. Classification accuracy of the optimised 1D-CNN for 32-QAM signals at different transmission distances.

The next experiment investigates the effect of the symbol interval (N). Four values, ($N = 10$), 20, 40, and 80, are evaluated using 32-QAM. As shown in Fig. 9, increasing the number of samples generally improves the available signal information. However, the improvement does not continue indefinitely. In particular, the ($N = 40$) configuration provides higher classification accuracy than ($N = 80$), while requiring a substantially smaller input representation. This observation matters because increasing the symbol interval increases both the amount of input data and the computational burden without necessarily adding useful information.

Among the four investigated configurations, $N = 40$ provides the highest classification performance in the present preliminary evaluation while requiring a substantially smaller input representation than $N = 80$. Therefore, $N = 40$ is selected for the subsequent experiments as the best-performing configuration among the tested values in this dataset. This selection should not be interpreted as a universally optimal symbol-segment length for VLC systems, since the preferred value of N may

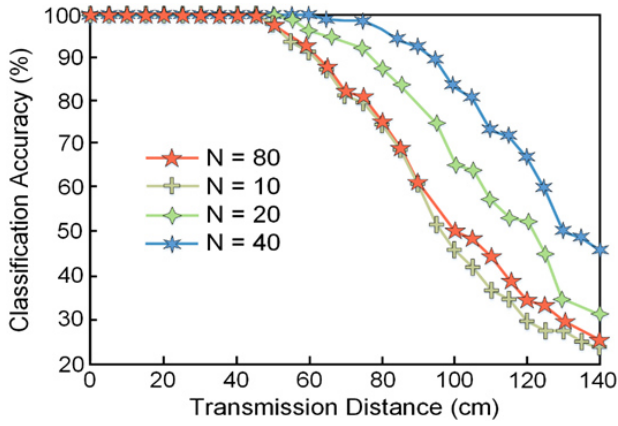


Fig. 9. 32-QAM classification accuracy obtained for different symbol interval (N).

depend on the sampling rate, waveform characteristics, channel conditions, and dataset. The result is also consistent with the behaviour reported for the original experimental dataset, where ($N=40$) provided a favourable trade-off between classification performance and data size. The influence of the training duration is subsequently investigated. The number of training epochs is varied from the initial training stage up to 200 epochs using 16-QAM and 32-QAM. As illustrated in Fig. 10, the classification accuracy increases during the initial training stage and gradually approaches a stable maximum as the number of epochs increases. Beyond approximately 200 epochs, no meaningful improvement is observed. Accordingly, 200 epochs are adopted for the subsequent experiments.

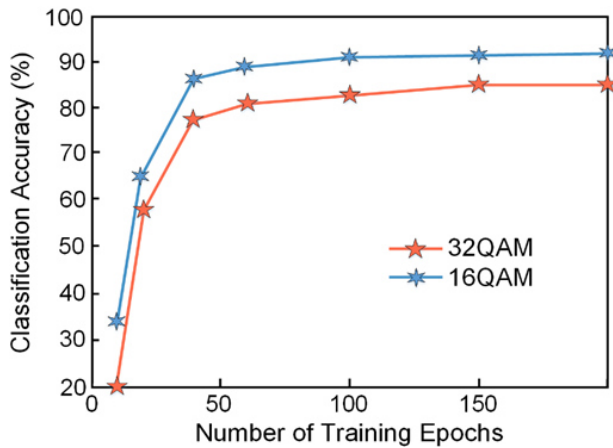


Fig. 10. Classification accuracy as a function of the number of training epochs for 16-QAM and 32-QAM signals.

The results of this subsection establish the principal network-development parameters used in the remaining experiments: a symbol interval of ($N=40$), a four-layer 1D-CNN architecture, and a maximum training duration of 200 epochs.

3.2. Effect of temporal context and selection of the $K=10$ configuration

Although the optimised 1D-CNN improves the classification performance, processing each signal segment independently provides the network with information from

only a single segment. To investigate whether additional neighbouring waveform information can improve classification, the temporal-context configurations defined in section 2 are evaluated using 32-QAM and $N=40$. The investigated configurations are $K=1, 3, 5, 8, 10$, and 12, where K denotes the number of consecutive signal segments jointly provided to the network. As shown in Fig. 11, symbol-demodulation accuracy increases as K increases from 1 to 10, while increasing to $K=12$ reduces accuracy. Thus, $K=10$ provides the highest classification accuracy among the temporal-context configurations evaluated in this study.

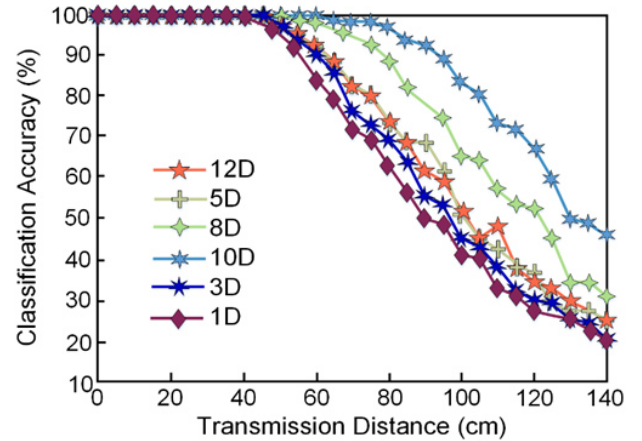


Fig. 11. Symbol-demodulation accuracy for 32-QAM at different temporal-context configurations $K=1, 3, 5, 8, 10, 12$ with $N=40$.

The observed improvement from $K=1$ to $K=10$ demonstrates that jointly processing multiple consecutive signal segments can improve classification performance for the investigated VLC dataset. However, the improvement cannot be uniquely attributed to multipath-induced temporal memory. Increasing K also increases the amount of waveform information available to the CNN, and the present study does not include a control experiment using shuffled or non-consecutive segments with the same total number of samples. Therefore, the results do not distinguish between the contribution of additional waveform information and the contribution of temporal correlations associated with the propagation channel. The present findings are consequently interpreted more conservatively as evidence that multiple consecutive signal segments provide useful information for classification. The reduction in performance at $K=12$ further indicates that increasing the temporal context beyond the best-performing tested configuration does not necessarily improve classification performance and may introduce redundant or less informative waveform information. Accordingly, $K=10$ is selected as the best-performing temporal-context configuration among those evaluated in this study, rather than as a universally optimal temporal window.

To provide a reference to the machine-learning-based VLC demodulation results reported in [23], the proposed CNN is compared with the AdaBoost-based result reported for 32-QAM in that study. Importantly, this comparison concerns 32-class symbol demodulation, rather than seven-class modulation-format recognition. In both cases, the

target is the transmitted symbol state within a fixed 32-QAM format, and the reported performance metric is symbol-demodulation accuracy. Therefore, this comparison serves only as a reference for 32-QAM symbol-demodulation performance and is not presented as a direct benchmark of modulation-format recognition. Figure 12 compares the proposed CNN with the reported AdaBoost result. The proposed CNN with temporal context achieves higher reported symbol-demodulation accuracy under the investigated conditions. However, the two studies do not employ an explicitly controlled, identical experimental protocol for all aspects of model training, data partitioning, and input construction. In particular, the proposed method jointly processes $K=10$ consecutive signal segments, whereas the AdaBoost method in [23] uses the signal representation defined in that study. Consequently, the comparison should be interpreted as a reference to previously reported 32-QAM symbol-demodulation performance rather than as a controlled attribution of the improvement to temporal context alone.

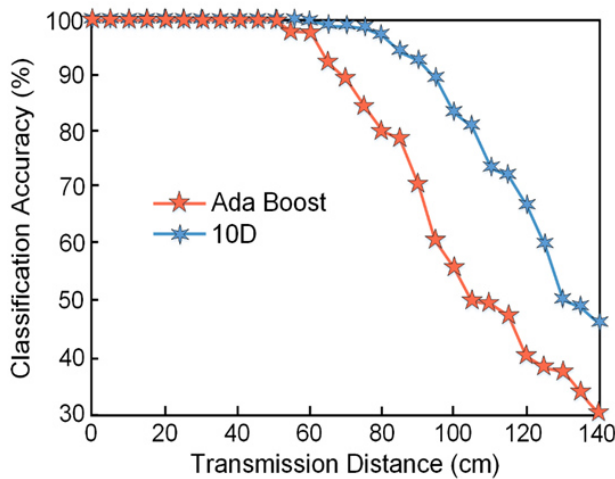


Fig. 12. Comparison of the proposed temporal-context CNN with the AdaBoost-based result reported in [23] for 32 QAM 32-class symbol demodulation. The comparison is presented as a reference to previously reported symbol-demodulation performance and not as a direct benchmark of seven-class modulation-format recognition.

The proposed temporal-context CNN provides higher symbol-demodulation accuracy than the AdaBoost result reported in [23] under the corresponding investigated conditions. This result indicates that the proposed CNN provides competitive 32-QAM symbol-demodulation performance relative to the previously reported AdaBoost approach. Nevertheless, this comparison is not a controlled ablation because the two approaches differ in model architecture and input representation, including the proposed method use of multiple consecutive signal segments. Therefore, the observed performance difference cannot be attributed specifically to temporal context, and the comparison is not direct evidence that temporal-context processing alone drives the improvement. For completeness, the architecture search performed for the $K=1$ baseline is summarised in Table 7, Table 8, and Table 9. These experiments demonstrate the sensitivity of the classification performance to both filter size and filter

number. The first configuration achieves a symbol-demodulation accuracy of 38%, whereas the second increases it to 48%. Further optimisation of the convolutional filters increases the symbol-demodulation accuracy to 62%.

Table 7.

Experimental CNN configuration I for 1D 32-QAM symbol demodulation.

Layers	Filter size	Number of filters
Conv-1	7×7	46
Conv-2	10×10	46
Conv-3	10×10	40
Conv-4	12×12	38
Symbol-demodulation accuracy	38%	–

Table 8.

Experimental CNN configuration II for 1D 32-QAM symbol demodulation.

Layers	Filter size	Number of filters
Conv-1	7×7	32
Conv-2	10×10	40
Conv-3	10×10	46
Conv-4	12×12	46
Symbol-demodulation accuracy	48%	–

Table 9.

Optimised CNN configuration for 1D 32-QAM symbol demodulation.

Layers	Filter size	Number of filters
Conv-1	12×12	38
Conv-2	10×10	40
Conv-3	10×10	46
Conv-4	7×7	46
Symbol-demodulation accuracy	62%	–

These experiments confirm that the CNN performance depends strongly on the relationship between convolutional receptive fields and the number of learned feature maps. The final architecture used in this study was therefore selected based on the experimentally observed classification performance rather than solely on architectural complexity.

3.3. Joint modulation and transmission-distance classification

After determining $N=40$ and the $K=10$ temporal-context configuration from the preliminary 32-QAM symbol-demodulation experiments, the proposed framework is evaluated using all seven investigated modulation formats: OOK, QPSK, 16-QAM, 32-QAM, 64-QAM, 128-QAM, and 256-QAM. Unlike the preliminary experiments, this stage evaluates the framework ability to distinguish among

different modulation formats and transmission distances. The corresponding modulation-format and transmission-distance results are therefore reported separately from the symbol-demodulation accuracies obtained during network development.

To make the overall classification results explicit, the final performance of the proposed joint network is evaluated on the held-out test subset described in section 2.4. The modulation-format classification accuracy of 99.5% and the transmission-distance classification accuracy of 98.7% correspond to the overall classification accuracies obtained for the seven modulation-format classes and 29 transmission-distance classes, respectively. For each task, the overall accuracy is defined as (8):

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100\% \quad (8)$$

where N_{correct} denotes the number of correctly classified test samples and N_{total} denotes the total number of evaluated test samples. Thus, the reported values represent overall sample-wise classification accuracy, in which each evaluated sample contributes equally to the final accuracy. They are not reported as macro-averaged class accuracies. Table 10 summarises the final overall results. Because the currently available result records retain the aggregate classification accuracies but do not retain the individual test-sample predictions and the corresponding complete sample-level evaluation records, the exact number of evaluated temporal-context test inputs and class-wise averaging statistics cannot be reconstructed retrospectively from the reported aggregate accuracies. Accordingly, no additional sample-level statistics are inferred from the aggregate results.

Table 10.

Overall classification performance of the proposed joint 1D-CNN on the held-out test subset.

Task	Number of classes	Overall test accuracy	Evaluation basis
Modulation-format classification	7	99.5%	Overall sample-wise accuracy
Transmission-distance classification	29	98.7%	Overall sample-wise accuracy

Class-level confusion analysis would provide additional information beyond the overall accuracies, particularly for distinguishing errors between neighbouring transmission-distance classes from errors involving more widely separated distances. A 7×7 confusion matrix would similarly allow examination of the class-wise behaviour of the seven modulation formats. However, the currently available results do not retain the individual predicted and ground-truth labels for the held-out test samples. The reported overall accuracies alone are insufficient to reconstruct either the 7×7 modulation-format confusion matrix or the 29×29 transmission-distance confusion matrix, because aggregate accuracy does not indicate which specific classes are involved in misclassifications. Therefore, class-level confusion matrices are not inferred from the reported accuracy values. For the same reason, the mean absolute distance error (MAE) cannot be obtained from the reported 98.7% classification accuracy alone. The

MAE requires the individual predicted and ground-truth distance labels and is calculated as (9):

$$\left[\text{MAE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |d_i^{\text{pred}} - d_i^{\text{true}}| \right], \quad (9)$$

where d_i^{pred} and d_i^{true} denote the predicted and ground-truth transmission distances for the i -th test sample. Because these sample-level predictions are not available in the retained results, an MAE value is not reported, but is instead estimated from the aggregate classification accuracy.

The network is trained using the full range of transmission distances and then evaluated over the individual distance intervals. Figure 13 shows the modulation-format classification results. A clear relationship emerges between modulation order, transmission distance, and classification accuracy. In general, the modulation-format classification accuracy decreases as the modulation order increases. This behaviour is expected because higher-order QAM formats contain a larger number of symbol states with smaller decision regions, making them more sensitive to waveform distortion, attenuation, and noise.

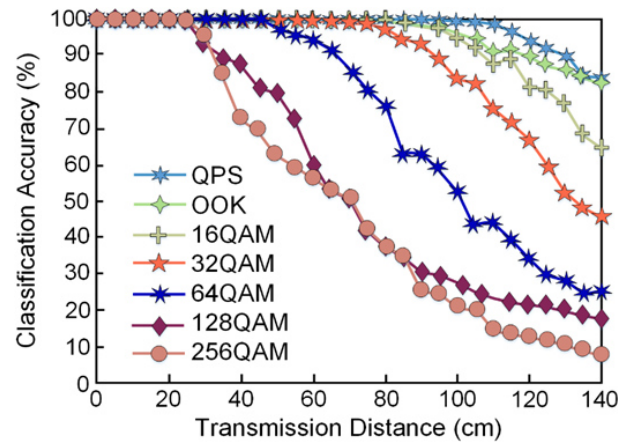


Fig. 13. Modulation-format classification accuracy vs. transmission distance for the seven investigated modulation formats using the selected $K = 10$ temporal-context configuration.

The results also show that the modulation-format classification accuracy tends to decrease as the transmission distance increases. The degradation becomes more pronounced for high-order QAM formats as the transmission distance increases. At longer distances, attenuation, noise, and channel-induced waveform distortion can reduce the distinguishability of the waveform characteristics used by the classifier. This effect is more evident for higher-order QAM formats, for which the received signal contains a larger number of closely spaced symbol states and consequently more complex waveform variations. In contrast, lower-order modulation formats such as OOK and QPSK retain comparatively higher classification accuracy over a larger distance range. The simultaneous influence of modulation order and transmission distance is particularly important for practical VLC systems. A classifier designed only for modulation recognition under a fixed channel condition may not provide reliable performance when the receiver position

changes. The proposed framework addresses this limitation by learning modulation-dependent features together with distance-dependent characteristics from the same received waveform. The corresponding transmission-distance classification results are presented in Fig. 14. Unlike modulation classification, the distance-classification accuracy does not exhibit a strictly monotonic behaviour with increasing distance. Instead, fluctuations are observed across the investigated intervals. This behaviour is expected because adjacent distance classes can produce relatively similar received waveforms, while variations in the experimental channel and multipath components can alter the received signal characteristics.

The results indicate that the received waveform contains sufficient distance-dependent information to discriminate among the 29 experimentally defined transmission-distance classes without a separate distance-estimation stage. In a VLC receiver, such a prediction could provide a coarse estimate of the receiver location along

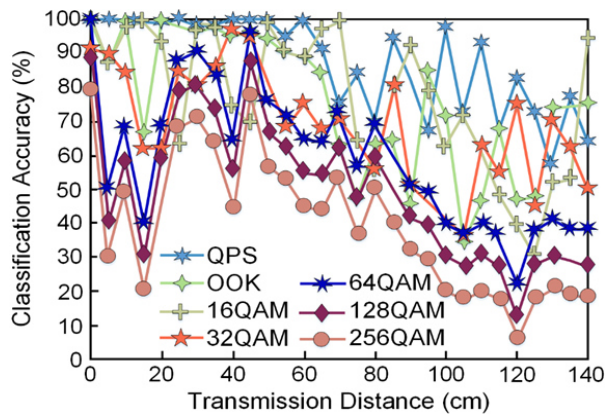


Fig. 14. Transmission-distance classification accuracy vs. distance for the investigated modulation formats using the $K = 10$ temporal-context configuration.

a calibrated link and could support link adaptation, configuration selection, or distance-dependent signal processing. Compared with estimating SNR or received power alone, waveform-based classification can potentially exploit multiple signal characteristics simultaneously, although the present study does not establish that it is superior to these conventional channel-state indicators.

The transmission-distance results should therefore be interpreted as discrete distance-class identification within the investigated experimental setup, rather than as general-purpose ranging. Distance-dependent waveform characteristics may also be affected by receiver alignment, orientation, ambient optical conditions, reflections, and hardware characteristics. Because the present dataset does not provide independent experimental configurations or acquisition sessions for validation, the reported 98.7% accuracy demonstrates discrimination among the tested laboratory positions under the investigated conditions and should not be interpreted as evidence of generalised distance estimation.

To place the proposed method in the context of recent research, Table 11 compares the proposed framework with representative machine-learning- and deep-learning-based VLC studies. Existing work has mainly focused on signal demodulation, LED non-linearity compensation, post-equalisation, or end-to-end transceiver optimisation. In particular, Ma *et al.* [23] introduced the experimentally acquired VLC dataset used in this study and evaluated CNN, DBN, and AdaBoost-based demodulation. However, they did not explicitly incorporate temporal dependencies between consecutive signal segments as a dedicated input representation.

The proposed network adopts a shared feature extractor followed by two task-specific classification heads, enabling modulation-format and transmission-distance inference within a single model. The primary motivation for this joint

Table 11.

Comparison of the proposed method with representative recent machine-learning-based VLC modulation-recognition studies.

Ref.	Data/VLC system	Modulation/ signal	Input representation	ML/DL method	Experimental validation	Main objective	Temporal information
[23]	Public real VLC dataset	OOK, QPSK, 16-, 32-, 64-, 128-, 256-QAM, 4-PPM	Signal-derived representation / image	CNN, DBN, AdaBoost	Yes	Physical-layer modulation demodulation	No explicit temporal-context modelling
[25]	NOMA-VLC system	NOMA signals	Received signal	CNN-based demodulator	Yes + simulation	Signal recovery and non-linear/linear distortion compensation	Implicit
[26]	Multi-CAP VLC	Multi-CAP signals	Received signal/features	Machine-learning clustering and predistortion	Yes	LED non-linearity compensation	No explicit temporal-context modelling
[27]	VLC system	VLC signals	Received signal/features	Probabilistic Bayesian learning	Yes	LED non-linearity estimation and compensation	No explicit temporal-context modelling
[28]	High-speed PAM VLC	PAM4, PAM8	Received waveform	Memory-controlled deep LSTM	Yes	Linear/non-linear impairment compensation and post-equalisation	Yes
[29]	Multi-coloured VLC	Multi-coloured VLC signals	Signal representation	Deep neural network	Yes	End-to-end VLC transceiver design	Implicit
[31]	High-speed VLC	Carrier-less amplitude and phase modulation	Received signal	Deep-learning-enhanced CAP	Simulation/ experimental evaluation	LED non-linearity and bandwidth limitation mitigation	Implicit
Our	Public experimentally acquired VLC dataset [23]	OOK, QPSK, 16-, 32-, 64-, 128-, 256-QAM	Raw temporal waveform	1D-CNN with temporal context	Yes	Joint modulation-format and transmission-distance classification	Yes; $K=1, 3, 5, 8, 10, 12$ with $K=10$ selected

architecture is simultaneous inference and shared feature extraction, not a claim of superior classification accuracy over independently trained single-task networks. Sharing the feature-extraction layers lets both classification tasks use the same learned representation, while the task-specific output heads preserve the distinct decision spaces of modulation format and transmission distance. This design is particularly relevant when both attributes are required from the same received waveform, since a single inference pipeline can produce the two predictions without deploying two completely independent feature-extraction networks. The present study does not include a controlled quantitative ablation against independently trained modulation-only and distance-only networks. Therefore, no claim is made that joint training necessarily improves the accuracy of either individual task relative to an independently optimised network. The high accuracies obtained for the two tasks, namely 99.5% for modulation-format classification and 98.7% for transmission-distance classification, demonstrate the effectiveness of the proposed joint architecture for simultaneous inference, but do not by themselves establish a performance advantage over separate networks. A systematic comparison under identical data partitions and training conditions is an important direction for future work. The AdaBoost results in [23] concern symbol demodulation under a known modulation format, rather than modulation-format recognition. Accordingly, these results are not used as a direct benchmark for the seven-class modulation-format classification task of the present study. Although reference [23] provides an important benchmark for machine-learning-based VLC demodulation, it does not explicitly investigate temporal-context processing of consecutive received signal segments. The present study uses the same experimentally acquired dataset while introducing a temporal-context representation that enables the network to exploit information distributed across neighbouring signal segments. Other studies have applied machine learning to specific VLC problems. The CNN-based demodulator in [25] focuses on signal recovery in NOMA-VLC, whereas references [26] and [27] employ machine-learning techniques for LED non-linearity estimation and compensation. Similarly, references [29] and [31] investigate deep-learning-based transceiver optimisation and mitigation of practical VLC impairments. These approaches confirm the effectiveness of deep learning in VLC but address objectives different from the joint classification problem considered here. Researchers have also exploited temporal dependencies in VLC. For example, reference [28] employs a memory-controlled deep LSTM network for post-equalisation of high-speed PAM-based VLC. In contrast, the proposed method introduces temporal information directly into a 1D-CNN by jointly processing consecutive received signal segments. The proposed approach therefore avoids recurrent architectures while retaining useful temporal information. A further distinction is that the proposed framework operates directly on the raw temporal waveform, without requiring constellation-image generation or other intermediate representations. The experimental results show that jointly processing multiple consecutive signal segments improves classification performance for the investigated dataset, with the $K=10$ configuration

providing the best performance among the tested configurations. The comparison with [23] further provides a reference for 32-QAM symbol-demodulation performance, but it should not be interpreted as a controlled comparison of modulation-format recognition or as direct evidence that temporal context alone causes the observed improvement. Overall, the proposed framework combines three characteristics not jointly addressed by the compared studies: raw-waveform processing, explicit temporal-context learning, and joint modulation-format and transmission-distance classification. Because datasets, channel conditions, modulation formats, and evaluation protocols differ across studies, Table 9 should be interpreted as a methodological comparison rather than a direct ranking of reported accuracy values.

4. Conclusions

This paper proposed a temporal-context 1D-CNN framework for experimentally acquired VLC signals, addressing symbol demodulation, modulation-format recognition, and transmission-distance classification. The network was first optimised using 32-QAM data. Because the modulation format was fixed in these preliminary experiments, the task was 32-QAM symbol demodulation using the 32 transmitted symbol states, rather than modulation-format recognition. Among the investigated configurations, $N=40$ and $K=10$ provided the best classification performance in the preliminary experiments and were therefore selected for the subsequent evaluation. The optimised framework was subsequently evaluated using seven modulation formats over transmission distances from 0 to 140 cm. The proposed method achieved 99.5% modulation-format classification accuracy and 98.7% transmission-distance classification accuracy. The results show that classification performance generally decreases with increasing modulation order and transmission distance, particularly for higher-order QAM formats, while OOK and QPSK maintain comparatively higher accuracy. The proposed approach integrates raw-waveform processing, temporal-context learning, joint modulation-format, and transmission-distance classification within a single 1D-CNN framework. The results demonstrate its potential for robust VLC signal characterisation, while further work on attention mechanisms, adaptive temporal windows, broader channel conditions, and real-time implementation could improve its generalisation and practical applicability.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The experimental VLC data used in this study were obtained from the publicly available dataset reported in [23]. This work uses a subset of this dataset after pre-processing and selecting the relevant modulation formats and transmission-distance classes.

References

- [1] Kim, T.-H. Analysis of optical communications, fiber optics, sensors and laser applications. *J. Mach. Comput.* **3**, 115–125 (2023). <https://doi.org/10.53759/7669/jmc202303012>
- [2] Wellbrock, G. A. *et al.* Explore benefits of distributed fiber optic sensing for optical network service providers. *J. Light. Technol.* **41**, 3758–3766 (2023). <https://doi.org/10.1109/JLT.2023.3263795>
- [3] Vikram, A., Patel, S. K., Chaturvedi, A., Alsalman, O. & Parmar, J. Detecting accurate parametric intrusions using optical fiber sensors for long-distance data communication system. *Opt. Fiber Technol.* **80**, 103453 (2023). <https://doi.org/10.1016/j.yofte.2023.103453>
- [4] Gopalan, A. *et al.* High performance efficiency of optical networking infrastructure characteristics with maximum speed of commercial fiber optic line cables employment. *J. Opt. Commun.* **45**, s2679–s2688 (2024). <https://doi.org/10.1515/joc-2023-0360>
- [5] Xu, K. & Yuan, C. A fault location analysis of optical fiber communication links in high altitude areas. *Electronics* **12**, 3728 (2023). <https://doi.org/10.3390/electronics12173728>
- [6] Tsonev, D., Videv, S. & Haas, H. Light fidelity (Li-Fi): Towards all-optical networking. *Proc. SPIE* **9007**, 900702 (2013). <https://doi.org/10.1117/12.2044649>
- [7] Chergui, S. & Abdessalam, S. Design and Realization of a Visible Light Communication System for Li-Fi Application. in *2020 1st International Conference on Communications, Control Systems and Signal Processing (CCSSP)* 30–35 (IEEE, 2020). <https://doi.org/10.1109/CCSSP49278.2020.9151780>
- [8] Nguyen, Q. D. & Nguyen, N. H. Mobile application for visible light communication systems: An approach for indoor positioning. *Photonics* **11**, 293 (2024). <https://doi.org/10.3390/photonics11040293>
- [9] Atiba, O. O. *Optical wireless and visible light communication techniques*. Master thesis, Tampere University (2023). <https://trepo.tuni.fi/handle/10024/148989>
- [10] Slaiman, I., Hamid, N. H. B. & Boon, T. T. Optical wireless communications through visible light LEDs: An overview. *Int. J. Eng. Res. Technol.* **3**, 1152–1156 (2014). www.ijert.org/research/optical-wireless-communications-through-visible-light-leds-an-overview-IJERTV3IS031111.pdf
- [11] Mapunda, G. A. *et al.* Indoor visible light communication: A tutorial and survey. *Wirel. Commun. Mob. Comput.* **2020**, 8881305 (2020). <https://doi.org/10.1155/2020/8881305>
- [12] Almadani, Y. *et al.* Visible light communications for industrial applications – challenges and potentials. *Electronics* **9**, 2157 (2020). <https://doi.org/10.3390/electronics9122157>
- [13] Manousiadis, P. P., Yoshida, K., Turnbull, G. A. & Samuel, I. D. W. Organic semiconductors for visible light communications. *Philos. Trans. R. Soc. A* **378**, 20190186 (2020). <https://doi.org/10.1098/rsta.2019.0186>
- [14] Fath, T. & Haas, H. Performance comparison of MIMO techniques for optical wireless communications in indoor environments. *IEEE Trans. Commun.* **61**, 733–742 (2013). <https://doi.org/10.1109/TCOMM.2012.120512.110578>
- [15] Liu, M., Song, T., Hu, J., Yang, J. & Gui, G. Deep learning-inspired message passing algorithm for efficient resource allocation in cognitive radio networks. *IEEE Trans. Veh. Technol.* **68**, 641–653 (2019). <https://doi.org/10.1109/TVT.2018.2883669>
- [16] Liu, X., Zhang, H., Long, K., Nallanathan, A. & Leung, V. C. Distributed unsupervised learning for interference management in integrated sensing and communication systems. *IEEE Trans. Wirel. Commun.* **22**, 9301–9312 (2023). <https://doi.org/10.1109/TWC.2023.3269815>
- [17] Hou, Z., She, C., Li, Y. & Vucetic, B. Mobility Prediction for Reducing End-to-End Delay in URLLC. in *Ultra-Reliable and Low-Latency Communications (URLLC) Theory and Practice: Advances in 5G and Beyond* (eds. Duong, T. *et al.*) 291–320 (Wiley, 2023). <https://doi.org/10.1002/9781119818366.ch10>
- [18] Abhaynarasimha, K. S., Mani, V. V. & Sellathurai, M. Mitigating nonlinear distortions of high-powered LEDs for VLC using deep neural networks. *Opt. Commun.* **550**, 129997 (2024). <https://doi.org/10.1016/j.optcom.2023.129997>
- [19] Li, Z. *et al.* Deep learning based end-to-end visible light communication with an in-band channel modeling strategy. *Opt. Express* **30**, 28905–28921 (2022). <https://doi.org/10.1364/OE.464277>
- [20] Shi, Y. *et al.* A Review of Machine Learning Based Techniques for Demodulation. in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)* 2292–2296 (IEEE, 2020). <https://doi.org/10.1109/ITAIC49862.2020.9339005>
- [21] Wang, J. *et al.* Learning-aided network association for hybrid indoor LiFi-WiFi systems. *IEEE Trans. Veh. Technol.* **67**, 3561–3574 (2018). <https://doi.org/10.1109/TVT.2017.2778345>
- [22] Saxena, V. N., Dwivedi, V. K. & Gupta, J. Machine learning in visible light communication system: A survey. *Wirel. Commun. Mob. Comput.* **2023**, 3950657 (2023). <https://doi.org/10.1155/2023/3950657>
- [23] Ma, S. *et al.* Signal demodulation with machine learning methods for physical layer visible light communications: Prototype platform, open dataset, and algorithms. *IEEE Access* **7**, 30588–30598 (2019). <https://doi.org/10.1109/ACCESS.2019.2903375>
- [24] Zeng, Y., Zhang, M., Han, F., Gong, Y. & Zhang, J. Spectrum analysis and convolutional neural network for automatic modulation recognition. *IEEE Wirel. Commun. Lett.* **8**, 929–932 (2019). <https://doi.org/10.1109/LWC.2019.2900247>
- [25] Lin, B., Lai, Q., Ghassemlooy, Z. & Tang, X. A machine learning based signal demodulator in NOMA-VLC. *J. Light. Technol.* **39**, 3081–3087 (2021). <https://doi.org/10.1109/JLT.2021.3058591>
- [26] Lu, X., Zhao, M., Qiao, L. & Chi, N. Non-Linear Compensation of Multi-CAP VLC System Employing Pre-Distortion Base on Clustering of Machine Learning. in *Optical Fiber Communication Conference M2K.1* (Optica Publishing Group, 2018). <https://doi.org/10.1364/OFC.2018.M2K.1>
- [27] Chen, C. *et al.* LED nonlinearity estimation and compensation in VLC systems using probabilistic Bayesian learning. *Appl. Sci.* **9**, 2711 (2019). <https://doi.org/10.3390/app9132711>
- [28] Lu, X. *et al.* Memory-controlled deep LSTM neural network post-equalizer used in high-speed PAM VLC system. *Opt. Express* **27**, 7822–7833 (2019). <https://doi.org/10.1364/OE.27.007822>
- [29] Lee, H., Lee, I. & Lee, S. H. Deep learning based transceiver design for multi-colored VLC systems. *Opt. Express* **26**, 6222–6238 (2018). <https://doi.org/10.1364/OE.26.006222>
- [30] Majeed, K. & Hranilovic, S. Passive indoor visible light positioning system using deep learning. *IEEE Internet Things J.* **8**, 14810–14821 (2021). <https://doi.org/10.1109/JIOT.2021.3072201>
- [31] Li, Z. Deep learning-enhanced carrier less amplitude and phase modulation for high-speed visible light communication: Mitigating LED nonlinearity and bandwidth limitations. *Ain Shams Eng. J.* **16**, 103534 (2025). <https://doi.org/10.1016/j.asej.2025.103534>
- [32] Zhou, S., Du, W., Li, C., Liu, S. & Li, R. Research progress on modulation format recognition technology for visible light communication. *Photonics* **12**, 512 (2025). <https://doi.org/10.3390/photonics12050512>