Applying Deep-Learning-Based Computer Vision to Wireless Communications: Methodologies, Opportunities, and Challenges

YU TIAN [1], GAOFENG PAN [1,2] (Senior Member, IEEE), AND MOHAMED-SLIM ALOUINI [1] (Fellow, IEEE)

[1]Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah
University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

[2]School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China

CORRESPONDING AUTHOR: G. PAN (e-mail: gaofeng.pan.cn@ieee.org)

This work was supported by King Abdullah University of Science and Technology (KAUST).

ABSTRACT Deep learning (DL) has seen great success in the computer vision (CV) field, and related techniques have been used in security, healthcare, remote sensing, and many other areas. As a parallel development, visual data has become universal in daily life, easily generated by ubiquitous low-cost cameras. Therefore, exploring DL-based CV may yield useful information about objects, such as their number, locations, distribution, motion, etc. Intuitively, DL-based CV can also facilitate and improve the designs of wireless communications, especially in dynamic network scenarios. However, so far, such work is rare in the literature. The primary purpose of this article, then, is to introduce ideas about applying DL-based CV in wireless communications to bring some novel degrees of freedom to both theoretical research and engineering applications. To illustrate how DL-based CV can be applied in wireless communications, an example of using a DL-based CV with a millimeter-wave (mmWave) system is given to realize optimal mmWave multiple-input and multiple-output (MIMO) beamforming in mobile scenarios. In this example, we propose a framework to predict future beam indices from previously observed beam indices and images of street views using ResNet, 3-dimensional ResNext, and a long short-term memory network. The experimental results show that our frameworks achieve much higher accuracy than the baseline method, and that visual data can significantly improve the performance of the MIMO beamforming system. Finally, we discuss the opportunities and challenges of applying DL-based CV in wireless communications.

INDEX TERMS Computer vision, deep learning, multiple-input and multiple-output, beamforming, beam tracking, long short-term memory, wireless communications.

I. INTRODUCTION

RECENTLY, deep learning (DL) has seen great success in the computer vision (CV) field. DL networks comprise networks such as deep neural networks, deep belief networks, recurrent neural networks (RNNs), and convolutional neural networks (CNNs). Many DL networks with various structures have emerged with the availability of large image and video datasets and high-speed graphic processing units (GPUs) [1]. DL networks can achieve success in CV because they discover and integrate low-/middle-/high-level features in images and leverage them to accomplish specific

tasks [2]. DL can easily fulfill CV applications with remarkably high performance, such as semantic segmentation, image classification, and object detection/recognition [1]. DL-based CV has therefore been widely utilized in public security, healthcare, and remote sensing, as such fields generate much visual data [3]. However, DL-based CV is rarely seen in the design and optimization of wireless communication systems in which the researchers mainly focus on the transmission quality of the information bits/packets, e.g., transmission rate, bit/packet error, traffic/user fairness, etc. via purely exploiting the information on the transmission

behaviors of radio frequency signals (e.g., the power, direction, phase, transmission duration, etc.), rather than making use of the geometry information of the surrounding space. Thus, such presented design and optimization of wireless communications cannot achieve the optimal performance with no doubts.

Nowadays, high-definition cameras are installed almost everywhere because of their low cost and small size. In some public areas, cameras have long existed for monitoring purposes. Therefore, visual data can easily be obtained in wireless communication systems in real life [4]. As useful information about *static system topology* (including terminals' numbers, positions, distances among themselves, etc.) and *dynamic system information* (including moving speed, direction, and changes in the number of the terminals) can be recognized, estimated, and extracted from these multi-medium data via DL-based CV techniques, new potential benefits can be exploited for wireless communications to aid system design/optimization, such as resource scheduling and allocations, algorithm design, and more.

Fig. 1 presents the framework of applying DL-based CV to wireless communications, the core idea of which is to explore the useful information obtained/forecasted by DL-based CV techniques to facilitate the design of wireless communications via DL-based/traditional optimization methods. In the following, we introduce some applications of DL-based CV in wireless systems in three aspects: the physical layer, medium access control (MAC) layer, and network layer.

1) In the physical layer of wireless communication systems, traditional methods usually first estimate the channel state by sending pilot signals from the transmitter to receivers [5]. Then according to the achieved channel state information (CSI), specific modulation, source encoding, channel encoding, and power control strategies can be selected to realize the optimal utilization of system resources (e.g., bandwidth and energy budgets). However, the CSI only contains amplitudes and phases information of the channel fading rather than the locations, number, and environmental information of the users which can be easily obtained from visual data by object detection and segmentation techniques in CV, leading to the fact that the real optimal system performance cannot be realized. However, with the aid of the more comprehensive users' information, dynamic modulation, encoding, and power control can be easily and optimally formulated and implemented. For example, in multiple-input and multiple-output (MIMO) beamforming communication systems, the direction and power of beams can be scheduled using the knowledge of users' locations and blocking cases in the visual data, which cannot be obtained via traditional methods.

2) In the MAC layer, like in cellular wireless networks, receiver-to-transmitter feedback information and cell-to-cell CSI is very important information to be utilized to allocate resources and to guarantee the quality of service in the traditional methods [6]. Thus, a long time delay may always exist when analyzing the feedback and CSI in crowded scenarios in which a huge number of users are served by the network. By jointly using these information and the density or distribution of users obtained from the visual data in the serving area of the BS, channel resources (including frequency bands, time slots, etc.) can be efficiently reserved and allocated to achieve optimal overall performance. For example, smart homes have various kinds of terminals such as smartphones, televisions, laptops, and other intelligent home appliances. As such, channel resources can be dynamically scheduled by considering the information obtained from the visual data, such as the number and locations of the users. Unlike traditional handover algorithms that adopt the measured fluctuation of received signal power to estimate the distance between the terminal and BS, the moving information including velocity and its variations can be fully estimated from visual data to accurately facilitate channel resource allocation in the handover process. This will be quite useful in fifth-generation wireless networks due to the shrinking sizes of the serving zones.

3) For the network layer, taking multi-hop transmission scenarios as an example, traditional routing algorithms are mostly running based on the length of the routing path estimated by the pilot and feedback signals which cannot reflect the timely location changes in mobile scenarios [7]. By exploiting system topology information from the visual data, novel routing algorithms can be designed to efficiently improve transmission performance, such as the end-to-end delivery delay, packet loss rate, jam rate, and system throughput. For another instance, wireless sensor networks have numerous sensors that can be deployed in target areas to monitor, gather, and transmit information about their surrounding environments. Then, the system topology information from visual data can be used to design multi-hop transmissions, which are required due to the inherent resource limitations and hardware constraints of the sensors.

In general, traditional algorithms adopted in wireless communication systems depend on traditional channel/network state estimation methods to grab the CSI and network state information which unavoidably suffer time delay and/or feed errors, resulting in low efficiency or even wrong decisions. Especially, it is hard or impossible to get accurate CSI or network state information in high dynamic network scenarios through traditional methods. Thanks to the inherent merits of DL-based CV techniques, the static and dynamic system information can be accurately and efficiently extracted from visual data, bringing vital benefits to the design and optimization of wireless communication systems.

In this context, this article introduces the methodologies, opportunities, and challenges of applying DL-based CV in wireless communications as an essential reference/guide for theoretical research and engineering applications.

The rest of this article is organized as follows. Section II overviews related work from two perspectives: datasets and applications. Section III presents an example of applying a DL-based CV to mmWave MIMO beamforming and
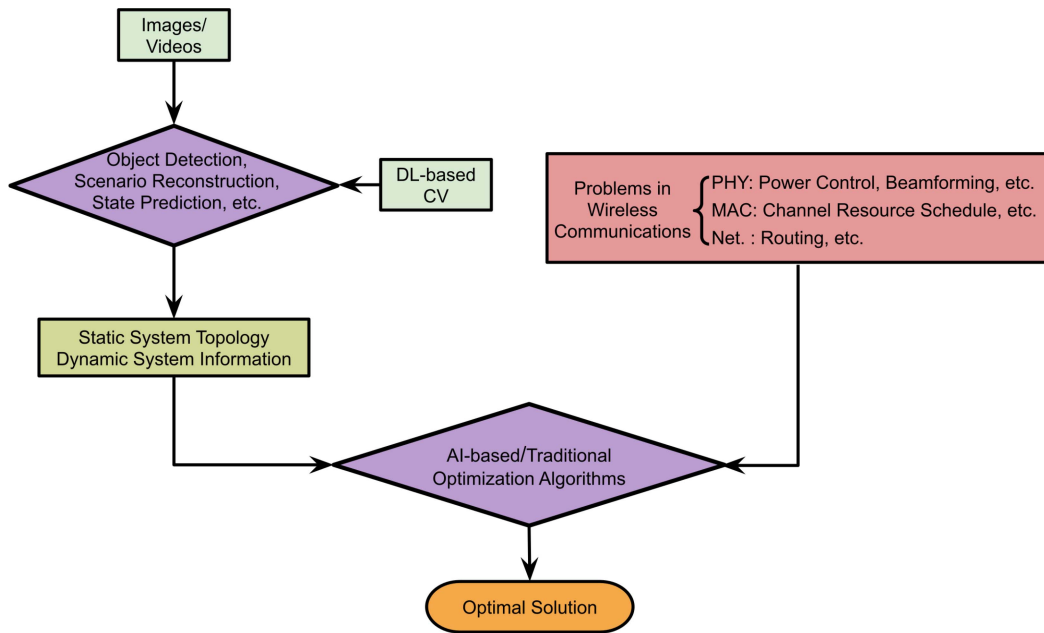
FIGURE 1. Framework of applying DL-based CV to wireless communications (PHY: Physical layer; MAC: MAC layer; Net.: Network layer; AI: artificial intelligence).

elaborates on the problem definition, framework architecture, pipeline, the results of the example, and practical application. Section IV introduces and discusses some challenges and open problems of applying DL-based CV to wireless communications. Finally, Section V concludes this article.

II. AN OVERVIEW OF RELATED WORK

Applying DL-based CV to wireless communications has two essential dimensions: datasets and applications. In the following, we give a brief overview of recent work in these two aspects.

a) *Datasets:* Building datasets is an essential step as DL is data-hungry. In [8], the authors proposed a parametric, systematic, scalable dataset framework called Vision-Wireless (ViWi). They utilized this framework to build the first-version dataset containing four scenarios with different camera distributions (co-located and distributed) and views (blocked and direct). These scenarios were based on a millimeter wave (mmWave) MIMO wireless communication system. Each scenario contained a set of images captured by the cameras and raw wireless data (signal departure/arrival angles, path gains, and channel impulse responses). Using the provided MATLAB script, they could view the user's location and channel information in each image from the raw wireless data. Later, the same authors built the second-version dataset called ViWi Vision-Aided Millimeter-Wave Beam Tracking (ViWi-BT) [9] and posted it for the machine learning competition at the IEEE International Conference on Communications (ICC) 2020. This dataset contains images captured by the co-located cameras and mmWave MIMO beam indices under a predefined codebook. Section III-D1 covers the details of this dataset. The authors of [10] introduced another dataset called Raymobtime which contains

ray-tracing, LIDAR, matrix channel, GPS, and image data in mmWave MIMO vehicle-to-infrastructure wireless communication systems. Notably, ray-tracing data provides path parameters such as received power, time of arrival, angle of departure, angle of arrival, line of a sight ray status, and ray phase while GPS user info data has line-of-sight (LOS) status, channel valid or not information, number of the TX in the vehicle, and the 3D coordinates. Beam selection and channel estimation challenges are held based on this dataset in International Telecommunication Union (ITU) Artificial Intelligence/Machine Learning in 5G Challenge. As both ViWi and Raymobtime datasets are versatile, except the challenges in ICC and ITU, many other interesting applications (e.g., blockage prediction, power prediction, angle estimation, LOS decision, etc.) can be explored according to their abundant data. In [11], a dataset consisting of depth image frames from recorded videos was built and can be applied in channel estimation tasks.

b) *Applications:* There are plenty of interesting applications of DL-based CV techniques designed to tackle problems in wireless communications. A framework to implement beam selection in mmWave communication systems by leveraging environmental information was presented by [4]. The authors used the images with different perspectives captured by one camera to construct a three-dimensional (3D) scene and generate corresponding point cloud data. They built a model based on 3D CNN to learn the wireless channel from the point cloud data and predict the optimal beam. Based on the first-version ViWi dataset, [12] proposed a modified ResNet18 model to conduct beam and blockage prediction from the images and channel information. Based on the second-version ViWi-BT dataset, the authors of [9] provided a baseline method based on Gated Recurrent

TABLE 1. Paper review.

Paper	Task	Network	Data type	Dataset public available?
[4]	Beam selection	3D CNN	Point clouds generated from camera images	No
[12]	Beam and blockage prediction	Modified ResNet18	Camera images	Yes
[9]	Beam prediction	GRU	Previous beams	Yes
[10]	Beam selection	CNN and DRL	Images generated from vehicle locations	Yes
[13]	Beam selection	CNN	LIDAR point clouds	Yes
[14]	Blockage prediction	CNN and RNN	Camera images and previous beams	Yes
[11]	Channel estimation	CNN	Depth images	Yes
[15]	Received power prediction	CNN and convLSTM network	Depth images	No
[16]	Handover decision	DRL	Camera images	No
[17]	Received power prediction	Multimodal split learning and convLSTM networks	Camera images and radio frequency signals	No

Units (GRUs) without the images, only the beam indices. They believe they can achieve better performance if they leverage both kinds of data. Based on the Raymobtime dataset, CNN and deep reinforcement learning (DRL) were utilized to select proper pair of beams for vehicles with images generated from GPS locations data in vehicle-to-infrastructure scenarios in [10]. The authors also compared DL-based methods with other traditional machine learning methods such as SVM, AdaBoost, decision tree, and random forest. The results showed that DL-based method has the best performance. In [13], two CNNs were proposed to conduct line-of-sight decision and beam selection by using LIDAR point cloud data in the Raymobtime dataset. In [14], authors proposed a neural network containing CNNs and an RNN-based recurrent prediction network to predict the dynamic link blockages using red, green, blue (RGB) images and beamforming vectors provided by the extended ViWi-BT dataset. In [11], authors developed a CNN-based framework called VVD to estimate the wireless communication channels only from only depth images in mmWave systems. In [15], a framework consisting of CNN and convolutional LSTM (convLSTM) network was presented to proactively predict the received power through depth images in mmWave networks and exhibited the highest accuracy compared with the random forest algorithm and a CNN-based method. In [16], a proactive handover management framework was proposed to make handover decisions by using camera images and DRL. In [17], a multimodal split learning method based on convLSTM networks was presented to predict mmWave received power through camera images and radio frequency signals while considering communication efficiency and privacy protection. All aforementioned papers are summarized in Table 1 for comparison purposes.

III. AN EXAMPLE OF APPLYING DL-BASED CV TO BEAMFORMING

A. PROBLEM DEFINITION

MmWave communication is a promising technique in the fifth-generation communication system, thanks to its broad available bandwidth and ultra-high data-transmitting rate [8], [9], [12]. MIMO and beamforming are widely used in mmWave communication systems and should be implemented in a large antenna array to achieve the required high
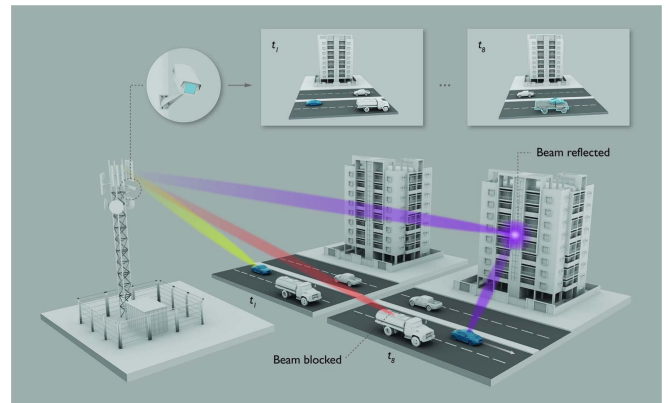

FIGURE 2. Scenario of applying DL-based CV to mmWave MIMO beamforming.

power gain and direction. The classic beamforming and beam tracking algorithms suffer a common disadvantage: complexity increases dramatically with the number of antennas, resulting in substantial computational overhead. DL-based CV is a promising candidate to address this overhead issue.

In this section, we give an example of applying the DL-based CV to mmWave MIMO beamforming. As defined in [9], the considered scenario contains 2 BSs located on the opposite sides of a street with a distance of 60 meters. As shown in Fig. 2, each BS forms a MIMO beam to serve a target user moving along the street. Therefore, the beam direction must be dynamically adjusted to catch the target mobile user. The target user may be blocked at some moments, such as t_8 in Fig. 2, and then the beam cannot directly reach the target user, while proper reflection from other objects, such as buildings and vehicles, must be designed. Meanwhile, three cameras installed at the BS capture RGB images of the whole street view to assist the beamforming process. The problem here is how to utilize the eight pairs of previously-observed consecutive beams and corresponding images to predict the future one, three, and five beams through DL model. Notably, these beams are represented as beam indices under the same predefined codebook.

A sequence containing the eight pairs of previously-observed images and corresponding beam indices for the uth user at the time instance t is given as

$$S_u[t] = \{(X_u[i], b_u[i])\}_{i=t-7}^{t}, \tag{1}$$
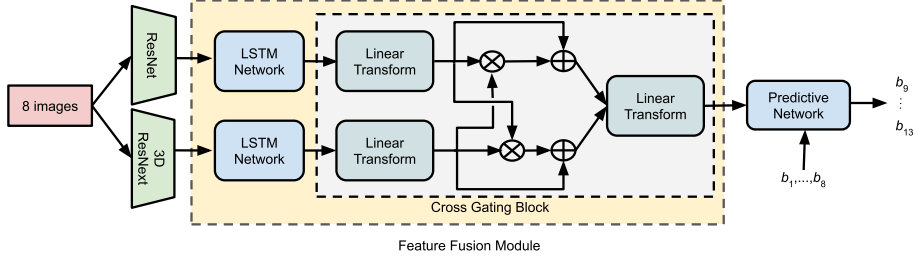
FIGURE 3. Architecture of our proposed framework.

where $X_u[i]$ is the RGB image taken at the *i*th time instance and $b_u[i]$ is the corresponding beam index.

Let $f_\Theta(S_u[t])$ be a prediction function of a DL model and $\hat{b}_u[t+n]$ ($n = 1, \ldots, 5$) be the predicted beam index at the time instance $t+n$. $f_\Theta(S_u[t])$ takes in the sequence $S_u[t]$ and outputs a predicted sequence $\{\hat{b}_u[t+n]\}_{n=1}^5$. Θ is a set of parameters of the DL model which is obtained by training the model with the training set. The training set consists of labelled sequences, i.e., $\mathcal{D} = \{(S_u[t], \{g_u[t+n]\}_{n=1}^5)\}_{u=1}^U$ where each pair consists of an observed sequence and five groundtruth future beam indices.

Equivalent to the defined problem, our goal is to get the prediction function which can maximize the joint success probability of all data samples in $\mathcal{D}$. The object function is expressed as

$$\max_{f_\Theta(S_u[t])} \prod_{u=1}^U \prod_{n=1}^5 \mathbf{Pr}\left\{ \hat{b}_u[t+n] = g_u[t+n] \Big| S_u[t] \right\}, \quad (2)$$

where each success probability only relies on its observed sequence $S_u[t]$.

B. FRAMEWORK ARCHITECTURE AND METHODS

We propose a DL network framework shown in Fig. 3 composed of ResNet [2], 3D ResNext [18], a feature-fusion module (FFM) [19], and a predictive network.

1) RESNET, RESNEXT AND 3D RESNEXT

ResNet consists of several residual blocks, as presented in Fig. 4. Each block contains two or more convolutional layers and superimposes its input to its output through identity mapping. It can efficiently address the vanishing gradient issue caused by the rising number of convolutional layers. If a specific number of such blocks are concatenated, as depicted in Fig. 4, ResNet is available to achieve as many as 152 layers.

Fig. 4 also presents the structure of the ResNext block [20], an improved version of the residual block, that adds a 'next' dimension, also called 'Cardinality'. It sums the outputs of *K* parallel convolutional layer paths that share the same topology and inherits the residual structure of the combination. As *K* diversities are achieved by *K* paths, this block can focus on more than one specific feature representations of the images.
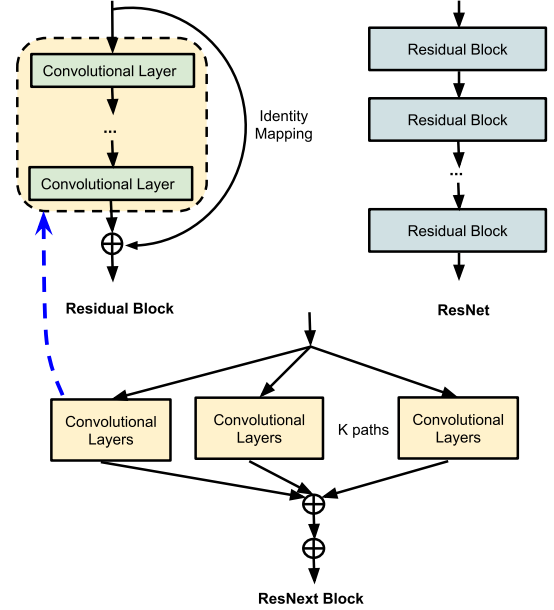
FIGURE 4. Structure of residual block, ResNet, and ResNext block.

In 3D ResNext, a similar structure can be observed but with 3D convolutional layers instead of two-dimensional (2D) ones. The 3D convolutional layer is designed to capture spatiotemporal 3D features from raw videos.

ResNet and 3D ResNext have been widely used as feature extractors for their powerful feature-representation abilities. If they are used in a DL network directly, however, the training time will become extremely long, and many computational resources will be occupied due to the large number of layers. Therefore, researchers commonly apply a pre-trained ResNet on the ImageNet dataset to extract visual features from images and a 3D ResNext on the Kinetics dataset to extract spatiotemporal features from videos [21]. These features are then fed to the DL network as inputs.

2) LONG SHORT-TERM MEMORY (LSTM) NETWORK

The LSTM [22] network is designed for the tasks that contain time-series data, such as prediction, speech recognition, text generation, etc. Hence, it is a suitable candidate for our predictive network. The network comprises several LSTM cells, as depicted in Fig. 5. Event (current state),
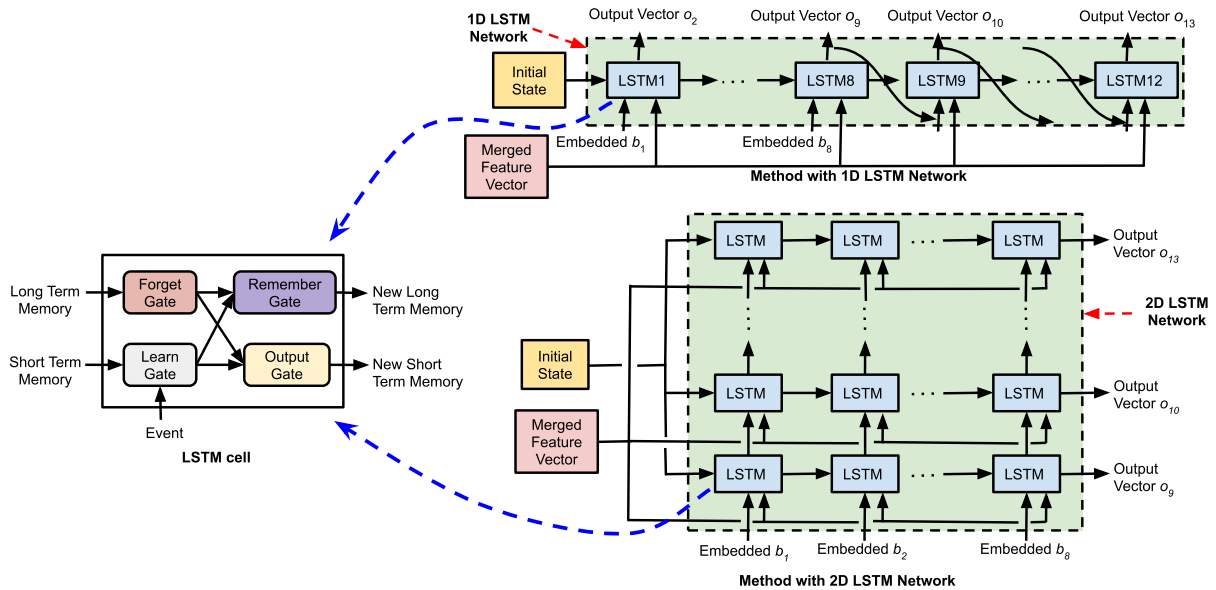
FIGURE 5. Structure of LSTM cell, method with 1D LSTM network, and method with 2D LSTM network.

previous long-term memory (cell state), and previous short-term memory (hidden state) are the inputs of an LSTM cell, in which learn, forget, remember, and output gates are employed to explore the information from the inputs. The LSTM cell outputs new long-term memory and short-term memory in which the latter is also regarded as a prediction.

When an LSTM cell is recursively utilized in a 1D array form, a 1D LSTM network is obtained, as presented in Fig. 5. At each moment, the cell and hidden states of the previous moment are used to generates the outputs of the current moment.

As shown in Fig. 5, a 2D LSTM network can be realized when the LSTM cell is recursively in a 2D mesh form [23]. Each LSTM cell utilizes the hidden and cell states from the two neighboring cells in the left and below positions in the mesh, and its states are delivered to its neighboring cells in the right and top positions. Obviously, the number of predictions is equal to the number of rows.

3) FEATURE-FUSION MODULE (FFM)

Fig. 3 shows the structure of the FFM which comprises two LSTM networks and a cross-gating block. The features from ResNet and 3D ResNext are aggregated by the LSTM networks and then high-level features are obtained. The cross-gating block can make full use of the related semantic information between these two kinds of features by multiplication and summation operations. Thus, the merged features can be obtained through a linear transformation.

C. PIPELINE OF OUR FRAMEWORK

In the pipeline of the considered DL network, eight consecutive images are inputted and utilized. As each is equivalent to a video clip, they contain motion information, which is helpful for the beam prediction. Combined with the visual

information from each image, location, motion, and blockage information can be extracted from these RGB images. The pre-trained 3D ResNext with 101 layers (3D ResNext101) is adopted to extract motion features and the pre-trained ResNet with 152 layers (ResNet152) is used to extract visual features. These features are then merged as a vector through FFM and sent to the predictive network.

As depicted in Fig. 5, there are three kinds of inputs in the predictive network, namely, initial state, embedded beam vectors, and merged feature vector. The initial state is set as a vector of all zeros. Embedding is mapping a constant (beam index) to a vector and can represent the relation between constants well. So the embedded vector is utilized to represent the beam index. In each LSTM cell, the embedded beam vector and the merged feature vector are firstly transformed to the same shape and then sum them up as the 'event'. According to the event, short and long term memories are obtained from the previous LSTM cell, and each cell predicts future 1 output vector whose index of the maximum element is the beam index. Notably, all the LSTM cells share the same merged features.

Based on the 1D and 2D LSTM networks introduced in Section III-B2, three methods are proposed and explained below.

1) METHOD WITH 1D LSTM NETWORK

When the predictive network in Fig. 3 is a 1D LSTM network, the first method is obtained, as presented in Fig. 5. The LSTM cell is recursive 12 times. The cell at the kth moment is denoted as the 'kth LSTM cell'.

As shown in Fig. 6, during the training process, the pipeline of our first method is the following.

Step 1: Eight consecutive images are fed to the pre-trained ResNet152 and 3D ResNext101 and then visual features and motion features are obtained;
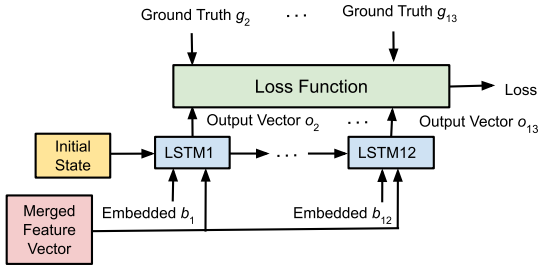
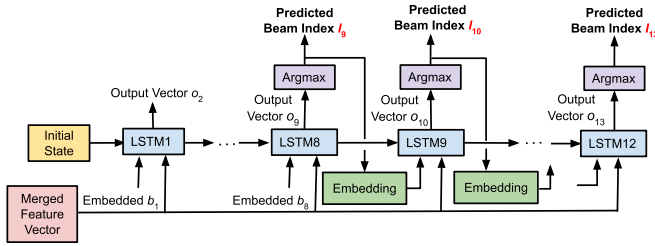**FIGURE 6.** Training procedure of method with 1D LSTM network.

FIGURE 7. Testing procedure of method with 1D LSTM network.

Step 2: These features from **step 1** are merged through the FFM;

Step 3: The output vector from the FFM is fed to each LSTM cell as an input;

Step 4: The embedded vectors of the first 12 beam indices go through the first to the last LSTM cells to update the hidden states and generate 12 output vectors;

Step 5: The 12 output vectors are used to calculate the training loss with the ground truth and train the network.

During the testing process, as we only have the first eight beam indices and images, the fourth step above is not applicable and is separated into two sub-steps as depicted in Fig. 7:

Substep 4.1: The embedded vectors of the first eight beam indices go through the first to seventh LSTM cells and update the hidden states;

Substep 4.2: The eighth to twelfth LSTM cells are used to predict the future beam indices which are obtained by acquiring the indices of the maximum element in these output vectors. Each cell is fed with the hidden state and the embedded beam index from the prediction of the previous LSTM.

The fifth step is skipped during testing.

2) METHOD WITH MODIFIED 1D LSTM NETWORK

In our first method, the training and testing procedures are different. Actually, the first method essentially aims to predict the next beam index as we utilize all the first 12 beam indices as inputs during the training. During the testing process, among the eighth to twelfth predicting beam indices, the previous one's correct prediction is important for the next prediction. To make training and testing processes consistent, we designed a modified version of the first method, in which the output vector of each of the last five LSTM cells undergoes a linear transformation module and is fed to the next cell as the embedded input. In this way, only the first eight beam indices are used as inputs, and the training and testing can be the same.

3) METHOD WITH 2D LSTM NETWORK

When we apply the 2D LSTM network to the predictive network, the third method can be realized as shown in Fig. 5. In this method, we need to input the initial state, the merged feature vector and embedded vectors of the first eight beam indices into the LSTM network, and then get five outputs vectors directly. The training process is the same as the testing one.

D. EXPERIMENT

In this section, we evaluate our three proposed methods on the ViWi-BT dataset. All the experiments are conducted in the framework of PyTorch on one NVIDIA V100 GPU.

1) DATASET

The VIWI-BT dataset contains a training set with 281,100 samples, a validation set with 120,468 samples, and a test set with 10,000 samples. There are 13 pairs of consecutive beam indices and corresponding images of street views in each sample of the training and validation sets. Furthermore, the first eight pairs are the observed beams for the target user and the sequence of the images where the target user appears. The last five pairs are groundtruth data containing the future beams and images of the same user. In this experiment, the first eight pairs serve as the inputs of the designed DL network to generate the predicted future five beam indices which are compared with the groundtruth ones.

2) IMPLEMENTATION DETAILS

We first use the pre-trained ResNet152 and 3D ResNext101 to extract 2048-dimensional visual and 8192-dimensional motion features from the first eight images of each sample. The merged features are embedded as a 463-dimensional vector and fed to the predictive LSTM network. There are a 512-dimensional hidden size and a 129-dimensional output vector in each LSTM cell. The training pipeline mentioned in Section III-C is then implemented to train the proposed network.

During the training, the designed DL network is optimized by the Adam optimizer. The learning rate is set as 4×10^{-4} at first and reduced by half every eight epochs. The batch size is set as 256. The cross-entropy loss is utilized for the loss function.

3) PERFORMANCE

Following the evaluation in [9], the performances of our proposed methods are evaluated on the validation set with the

TABLE 2. Performance of top-1 accuracy and exponential decay scores.

	Top-1 Accuracy			Exponential Decay Score			Running Time (s) for '1 future beam' prediction
	1 future beam	3 future beams	5 future beams	1 future beam	3 future beams	5 future beams	
Method with 1D LSTM Network	0.9170	0.7719	0.6448	0.9238	0.8206	0.7356	0.42+0.016
Method with Modified 1D LSTM Network	0.8887	0.6260	0.4800	0.8974	0.7137	0.6129	0.42+0.024
Method with 2D LSTM Network	0.8704	0.5893	0.4503	0.8803	0.6857	0.5877	0.42+0.064
Baseline Method in [9]	0.85	0.60	0.50	0.86	0.68	0.60	0.0016

same metrics, which are the top-1 accuracy and exponential decay score.

As defined in [9], the top-1 accuracy of n future beams is expressed as

$$Acc_{\text{top-1}}^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\left\{ \hat{\mathbf{g}}_i^{(n)} = \mathbf{g}_i^{(n)} \right\}, \tag{3}$$

where M is the number of the samples in the validation set, $\mathbb{1}\{\cdot\}$ denotes the indicator function, and $\hat{\mathbf{g}}_i^{(n)}$ and $\mathbf{g}_i^{(n)}$ represent the predictive and groundtruth beam indices vectors of the ith sample with the length n.

The exponential decay score of n future beams is given as

$$\text{score}^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \exp\left(-\frac{\left\| \hat{\mathbf{g}}_i^{(n)} - \mathbf{g}_i^{(n)} \right\|_1}{n\sigma} \right), \tag{4}$$

where $\sigma = 0.5$ is a penalization parameter.

Table 2 lists our results, in which the baseline method in [9] is considered for comparison purposes. In the baseline method, the authors simply leveraged the beam-index data and ignored image data.

From the top-1 accuracy, we can see that our proposed method with the 1D LSTM network outperforms the baseline method in [9]. The method with a modified 1D LSTM network is better than the baseline method on '1 future beam' and '3 future beams'. The method with only the 2D LSTM network performs better than the baseline method on '1 future beam'.

For the exponential decay scores, the designed methods with the 1D LSTM network and modified 1D LSTM network absolutely outperformed the baseline method. The method with the 2D LSTM network is better than the baseline on '1 future beam' and '3 future beams' but a little worse on '5 future beams'.

Our proposed methods outperform the baseline method on predicting '1 future beam'. Because the location, blockage, and speed information of the target user is extracted from the images and represented as motion and visual features to assist the prediction, and advanced LSTM networks are leveraged as the predictive networks. Among the three proposed methods, the method with the 1D LSTM network shows the best beam prediction performance in the target mobile scenarios. Compared with this method, the other two exhibit extra linear transformation modules or more LSTM cells in their predictive networks and need to be trained by more data. Therefore, performance degradation occurring on

the predictions of '3 future beams' and '5 future beams' are caused by the small size of the training dataset.

The computational complexity here is measured by the running time of '1 future beam' prediction which exhibits the best performance and is more potential to be implemented in the practical wireless communication systems. The running time of our methods consists of the execution time of feature extraction, FFM, and predictive network. It takes 0.42 seconds for the pre-trained 3D ResNext101 and ResNet152 to extract features from each set of eight images. Our method with 2D LSTM network exhibits the longest average running time due to its most complex structure shown in Fig. 5. The baseline method runs for the shortest time as it utilizes simple GRUs as the predictive network and abandons the image data. The method with 1D LSTM network shows the best predictive performance but a moderate prediction time, 0.016 seconds.

The feature extraction takes a little long time which will cause latency issues but can be mitigated by employing more efficient CNNs in future work. From [24, Table 1], we can see that the TSN [25] takes 15.5 ms and the ECO [26] takes 17.4 ms on NVIDIA Tesla P100 GPU to extract 3D spatiotemporal features from eight images. In practical application, only one new image is captured at each time instance, and the previously extracted spatiotemporal features can be merged with the new image to reduce the latency. Similarly, for the 2D feature extraction, only the new image needs to be processed at each time instance. In [27, Table 2] shows that the Darknet-19 can achieve a speed of 171 frames per second (5.85 ms per image) to extract 2D features on NVIDIA Titan X GPU. 2D and 3D features extractions can be conducted simultaneously which means that the longer latter determines the whole feature extraction time. By jointly using TSN and Darknet-19, the running time of the method with 1D LSTM network can be reduced to less than $15.5 + 16 = 31.5$ ms on V100 GPU which is more powerful than P100 and Titan X. Besides, in our experiments, we built the LSTM cell by ourselves. If we use the basic building block of LSTM in PyTorch, the comparable running time of the predictive network can be obtained with the baseline method. In engineering practice, even shorter latency can be achieved by utilizing GPU computing in the whole process and designing a more efficient framework (platform).

4) PRACTICAL APPLICATION

In practical scenarios, RGB images and their corresponding beam indices can be obtained by cameras installed on the

BS and the classic beamforming algorithm. After obtaining sufficient data for the training set, our proposed network will be pre-trained on these data and then run in the processors of the BS for beamforming. At the beginning of the serving time, the first eight beam indices can be estimated by the classic beamforming algorithm. Then the eight pairs of images and beam indices are sent to the processors for future beam predictions. Notably, after the first eight beams, the following beams will be predicted by using previously-obtained images and beam indices. These predicted beam indices and their corresponding images can be added to the training set to enlarge the dataset and improve the performance.

IV. CHALLENGES AND OPEN PROBLEMS
Although the previous sections elaborated on leveraging CV to tackle the mmWave beamforming problem, some challenges and open problems still exist in the front way of applying DL-based CV technologies in wireless communications.

A. BUILDING DATASETS
As DL is immensely data-hungry, a large dataset can guarantee the successful application of DL-based CV techniques to wireless communications. A qualified dataset in CV usually includes more than 10,000s of samples. For example, there are more than 14 million images in ImageNet, 60,000 images in Cifar-10, and roughly 650,000 video clips in Kinetics. It takes much time, money, and labor to generate such a huge amount of visual data. However, building a qualified dataset, which should be comprehensive and exhibit a balanced diversity of data, is still far from accomplished. These data should be able to represent all possibilities in the corresponding problem, and the amounts of different kinds of data can not make such a difference. Usually, a training set, a validation set, and a test set comprise a dataset. These three sets should be homogeneous and not overlapping, so randomly sampling them from a shuffled data pool is a better way to obtain the three sets. These data should be well-organized and easily manipulated. Therefore, the hardest work in DL is to build a satisfactory dataset.

B. SELECTING CV TECHNIQUES
Many state-of-the-art DL techniques have been proved efficient and powerful in CV, such as reinforcement learning, encoder-decoder architecture, generative adversarial networks (GANs), Transformer, graph convolutional networks (GCNs), etc. Reinforcement learning has been widely applied in tackling optimization problems in wireless communications [28]. GCNs can be leveraged to address network-related issues [29], and encoder-decoder architecture is widely used in semantic segmentation and sequence-to-sequence tasks. The GAN is an immensely powerful CNN to learn the statistics of training data and has been widely used to improve the performance of other DL networks in CV [1]. Transformer built on the attention mechanisms is a kind of the encoder-decoder architecture that
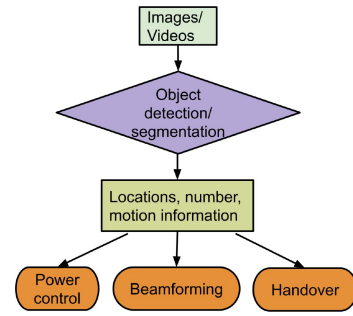

FIGURE 8. Pipeline of applying DL-based CV to cellular networks.
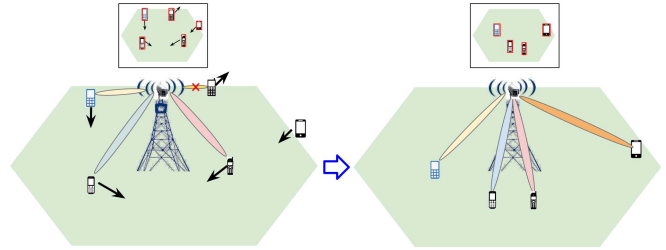

FIGURE 9. An example of applying DL-based CV to cellular networks.

can handle unordered sequences of data [30]. Much CV research has shown that if these techniques are jointly applied to make full use of the visual data, better results can be obtained [19], [21], [31].

Thus, a single proper CV technique or an adequate combination of several CV techniques are required to handle a specific problem in wireless systems. In the example given in Section III-B, we combined ResNet, 3D ResNext, and an LSTM network to achieve the required performance. Finding proper, efficient CV techniques thus remains an open problem.

C. OPEN PROBLEMS IN VISION-AIDED WIRELESS COMMUNICATIONS
As many kinds of cameras and LIDARs operate in real life, an enormous amount of visual data can be obtained through them, for more accurate motion and position information in the terminals that can be recognized, analyzed, and extracted from these multimedia data, which can also be explored to design and optimize wireless communications. Thus, some open problems in wireless communication scenarios are introduced and discussed as follows.

(1) *Cellular Networks:* As shown in Fig. 8, visual data obtained at the BS in a cellular network may contain the locations, number, and motion information of the terminals in the open area. This information can be used for the BS to adjust its transmitting power and beam direction to save power consumption and reduce interference. Fig. 9 presents a real-life example: the motion information of the users in the coverage area of a BS can be utilized to forecast the future positions of these terminals and judge whether/when a terminal goes out or comes into its serving area. Then, transmit power and beam can be accurately assigned for
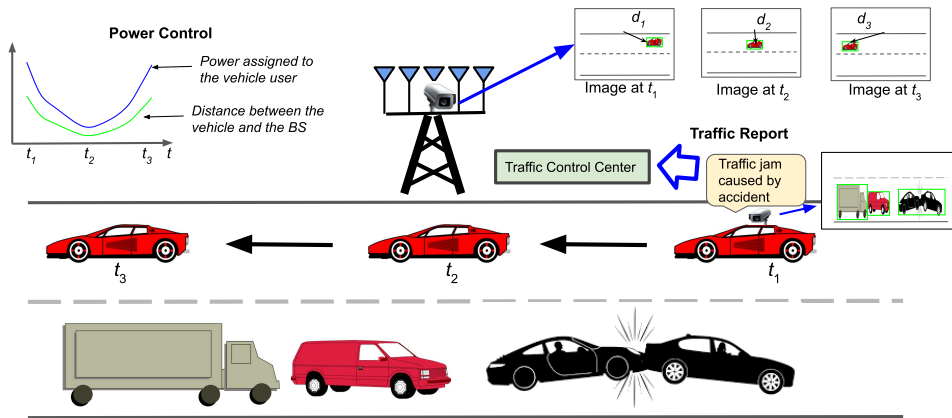
FIGURE 10. An example of applying DL-based CV to vehicle-to-everything communications.
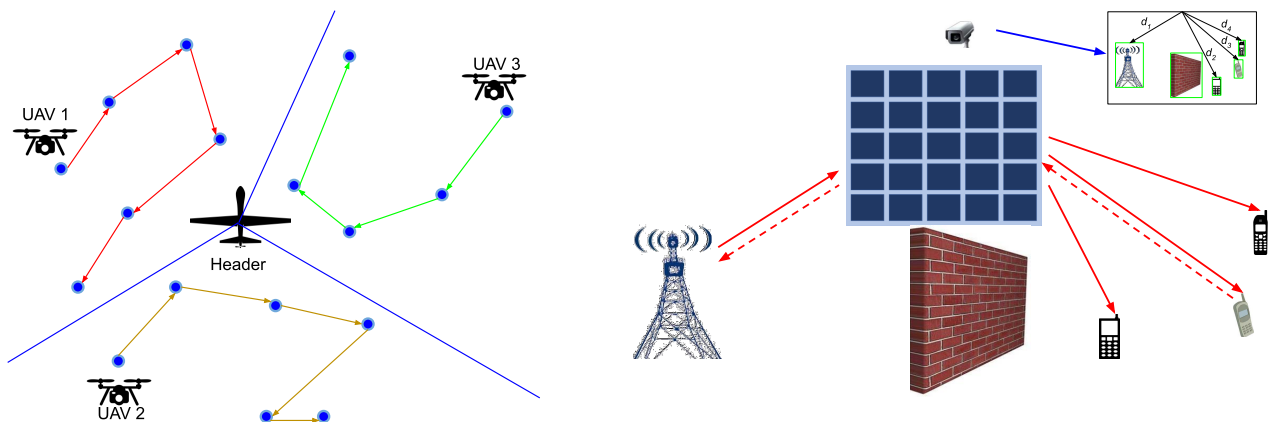

FIGURE 11. An example of applying DL-based CV to UAV-ground communications.

FIGURE 12. An example of applying DL-based CV to IRS system.

these users who still stay in the coverage area, and channel resource allocation can be set up for the handover process to improve the utilization efficiency of the system resource.

(2) *Vehicle-to-Everything Communications:* Visual data captured by one vehicle can reveal its environments, such as traffic conditions, which can be used to set up links with neighboring terminals, access points, and vehicles. Therefore, traffic schedules and jam/accident alarms can be conducted for improved road safety, traffic efficiency, and energy savings. As depicted in Fig. 10, the BS located on the side of the road utilizes the visual information to estimate the distance of a vehicle terminal from it and adjust its transmit power accordingly for power saving and interference deduction purposes. Moreover, the images or videos captured by the cameras assembled on the vehicles can detect the traffic jam or accident, and then forwards the observed traffic information to the traffic control center for alarm and future traffic schedule.

(3) *Unmanned Aerial Vehicle (UAV)-Ground Communications:* When a UAV serves as an aerial BS, visual data captured by the UAV can be used to identify the locations and distribution of ground terminals, which can be utilized in power allocation, route/trajectory

planning, etc. Moreover, when a ground BS communicates with several UAVs, visual data captured by the ground terminal can be used to define the serving range, allocate the channels/power, and so forth. Fig. 11 illustrates an example in which a group of UAVs communicates with a set of ground terminals. The header UAV first takes an image of the whole area and detects all the terminals. Then the serving area is divided into several subareas. Each UAV serves one specific subarea and designs a route schedule according to the location information of these ground terminals obtained from the images.

(4) *Smart Cities:* Visual data captured by satellites or airborne crafts can be applied to recognize and analyze the user's distribution and schedule power budget/serving ranges to achieve optimal energy efficiency.

(5) *Intelligent Reflecting Surfaces (IRSs):* Usually, implementing channel estimation and achieving network state information at an IRS is impossible because there is no comparable calculation capacity and no radio frequency (RF) signal transmitting or receiving capabilities at the IRS. Fortunately, DL-based CV can offer useful information to compensate for this gap. Thus, a proper control matrix can be optimally designed to accurately reflect the incident signals to the target destination by utilizing the visual data

captured by the camera installed on the IRS, which includes the locations, distances, and the number of terminals shown in Fig. 12.

V. CONCLUSION

This article mainly presented the methodologies, opportunities, and challenges of applying DL-based CV to wireless communications. First, we discussed the feasibility of applying a DL-based CV in physical, MAC, and network layers in wireless communication systems. Second, we overviewed related datasets and work. Third, we gave an example of applying a DL-based CV to a mmWave MIMO beamforming system. In this example, previously observed images and beam indices were leveraged to predict future beam indices using ResNet, 3D ResNext, and an LSTM network. The experimental results show that visual data can significantly improve the accuracy of beam prediction. Finally, challenges and possible research directions were discussed. We hope this work stimulates future research innovations and fruitful results.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[3] V. Bharti, B. Biswas, and K. K. Shukla, "Recent trends in nature inspired computation with applications to deep learning," in *Proc. 10th Int. Conf. Cloud Comput. Data Sci. Eng. (Confluence)*, Noida, India, Jan. 2020, pp. 294–299.

[4] W. Xu, F. Gao, S. Jin, and A. Alkhateeb, "3D scene based beam selection for mmWave communications," 2019. [Online]. Available: arXiv:1911.08409.

[5] S. Zhou and G. B. Giannakis, "Adaptive modulation for multiantenna transmissions with channel mean feedback," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1626–1636, Sep. 2004.

[6] D. Gesbert, S. G. Kiani, A. Gjendemsjo, and G. E. Oien, "Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks," *Proc. IEEE*, vol. 95, no. 12, pp. 2393–2409, Dec. 2007.

[7] C. Jung, K. Kim, J. Seo, B. N. Silva, and K. Han, "Topology configuration and multihop routing protocol for Bluetooth low energy networks," *IEEE Access*, vol. 5, pp. 9587–9598, 2017.

[8] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, "ViWi: A deep learning dataset framework for vision-aided wireless communications," 2019. [Online]. Available: arXiv:1911.06257.

[9] M. Alrabeiah, J. Booth, A. Hredzak, and A. Alkhateeb, "ViWi vision-aided mmWave beam tracking: Dataset, task, and baseline solutions," 2020. [Online]. Available: arXiv:2002.02445.

[10] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang, and R. W. Heath, "5G MIMO data for machine learning: Application to beam-selection using deep learning," in *Proc. Inf. Theory Appl. Workshop (ITA)*, 2018, pp. 1–9.

[11] S. Ayvaşık, H. M. Gürsu, and W. Kellerer, "Veni Vidi Dixi: Reliable wireless communication with depth images," in *Proc. 15th Int. Conf. Emerg. Netw. Exp. Technol.*, 2019, pp. 172–185.

[12] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision aided beam and blockage prediction," 2019. [Online]. Available: arXiv:1911.06255.

[13] A. Klautau, N. González-Prelcic, and R. W. Heath, "LIDAR data for deep learning-based mmWave beam-selection," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 909–912, Jun. 2019.

[14] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided dynamic blockage prediction for 6G wireless communication networks," 2020. [Online]. Available: arXiv:2006.09902.

[15] T. Nishio *et al.*, "Proactive received power prediction using machine learning and depth images for mmWave networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2413–2427, Nov. 2019.

[16] Y. Koda, K. Nakashima, K. Yamamoto, T. Nishio, and M. Morikura, "Handover management for mmWave networks with proactive performance prediction using camera images and deep reinforcement learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 802–816, Jun. 2020.

[17] Y. Koda *et al.*, "Communication-efficient multimodal split learning for mmWave received power prediction," *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1284–1288, Jun. 2020.

[18] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6546–6555.

[19] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, "Controllable video captioning with POS sequence guidance based on gated fusion network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct./Nov. 2019, pp. 2641–2650.

[20] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 1492–1500.

[21] J. S. Park, M. Rohrbach, T. Darrell, and A. Rohrbach, "Adversarial inference for multi-sentence video description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 6598–6608.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] P. Bahar, C. Brix, and H. Ney, "Towards two-dimensional sequence to sequence model in neural machine translation," 2018. [Online]. Available: arXiv:1810.03975.

[24] Z. Liu *et al.*, "Teinet: Towards an efficient architecture for video recognition," in *Proc. AAAI*, 2020, pp. 11669–11676.

[25] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 20–36.

[26] M. Zolfaghari, K. Singh, and T. Brox, "ECO: Efficient convolutional network for online video understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 695–712.

[27] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018. [Online]. Available: http://arxiv.org/abs/1804.02767.

[28] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[29] K. Rusek, J. Suárez-Varela, A. Mestres, P. Barlet-Ros, and A. Cabellos-Aparicio, "Unveiling the potential of graph neural networks for network modeling and optimization in SDN," in *Proc. ACM Symp. SDN Res.*, San Jose, CA, USA, Apr. 2019, pp. 140–151.

[30] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[31] Y. Hua, R. Li, Z. Zhao, X. Chen, and H. Zhang, "GAN-powered deep distributional reinforcement learning for resource management in network slicing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 334–349, Feb. 2020.

YU TIAN received the B.Sc. degree in communication engineering from the Harbin Institute of Technology in 2013, and the M.Sc degree from the King Abdullah University of Science and Technology, Saudi Arabia, in 2019, where he is currently pursuing the Ph.D. degree. His current research interests include deep learning and performance analysis of wireless communication systems.

GAOFENG PAN (Senior Member, IEEE) received the B.Sc. degree in communication engineering from Zhengzhou University, Zhengzhou, China, in 2005, and the Ph.D. degree in communication and information systems from Southwest Jiaotong University, Chengdu, China, in 2011. Since August 2019, he has been a Visiting Researcher with the Communication Theory Lab, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. He has also been a Professor with School of Cyberspace Science and Technology, Beijing Institute of Technology, China, since November 2020. His research interests include communications theory, signal processing, and protocol design.

MOHAMED-SLIM ALOUINI (Fellow, IEEE) was born in Tunis, Tunisia. He received the Ph.D. degree in electrical engineering from the California Institute of Technology, Pasadena, CA, USA, in 1998. He served as a Faculty Member with the University of Minnesota, Minneapolis, MN, USA, and then with the Texas A&M University at Qatar, Education City, Doha, Qatar. In 2009, he joined King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, as a Professor of Electrical Engineering. His current research interests include modeling, design, and performance analysis of wireless communication systems.