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

Yu Tian, Gaofeng Pan, *Senior Member, IEEE*, and Mohamed-Slim Alouini, *Fellow, IEEE*

***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 fields. 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, MIMO, 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. It comprises networks such as deep neural networks, deep belief networks, recurrent neural networks (RNNs), and convolutional neural networks (CNNs). Many DL networks have emerged with the availability of large image and video datasets and high-speed graphic processing units (GPUs) [1]. DL networks 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 generated [3]. However, DL-based CV is rarely seen in wireless communication in which one-dimensional temporal wireless data prevail.

Manuscript received**, 2020; revised **, 2020; accepted **, 2020. The associate editor coordinating the review of this paper and approving it for publication was ***. (Corresponding author: Gaofeng Pan.)

Authors are with Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.

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.

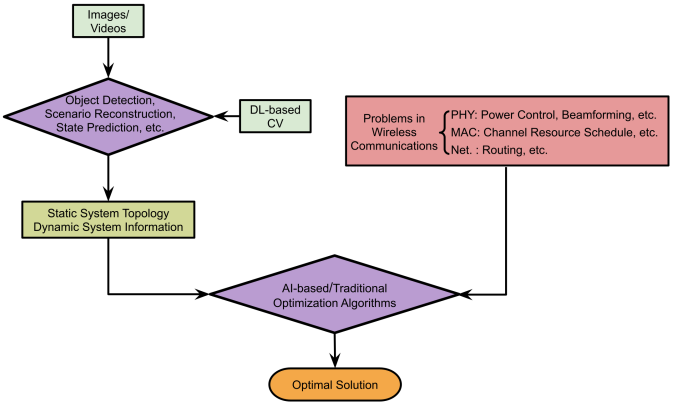

Fig. 1. Framework of applying DL-based CV to wireless communications (PHY: Physical layer; MAC: MAC layer; Net.: Network layer; AI: artificial intelligence)

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, wireless communication systems can leverage object detection and segmentation techniques in CV to obtain the locations, number, and environmental information

about users from visual data. With the aid of this information, 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). In this way, dynamic modulation, encoding, and power control can be easily 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.

2) In the MAC layer, using the density or distribution of users obtained from the visual data in the serving area of the base station (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 location 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, another example is moving information including velocity and its variations, which can be fully estimated from visual data to accurately facilitate channel resource allocation in the handover process. This can be quite useful in fifth-generation wireless networks due to the shrinking sizes of the serving zones.

3) For the network layer, in multi-hop transmission scenarios, novel routing algorithms can be designed to improve transmission performance, such as end-to-end delivery delay, packet loss rate, jam rate, and system throughput, by exploiting system topology information from the visual data. For 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, 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 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 elaborates on a problem definition, framework architecture, pipeline, and the results of the example. Section IV introduces and discusses some challenges and open problems of applying DL-based CV to wireless communications. Finally, Section V concludes the 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: DL is data-hungry, so building datasets is an essential step. In [5], the authors proposed a parametric, systematic, scalable dataset framework called Vision-Wireless (ViWi). They utilized this framework to build the first-version dataset with 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) [6] and posted it for the machine learning competition at the IEEE International Conference on Communications 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.

b) Applications: Interesting applications exist to tackle beamforming problems. A framework to implement beam selection in mmWave communication systems by leveraging environmental information was presented by [4]. They 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 and predict the optimal beam. Based on the first-version ViWi dataset, [7] 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 [6] provided a baseline method without the images, only the beam indices. They believe they can achieve better performance if they leverage both kinds of data.

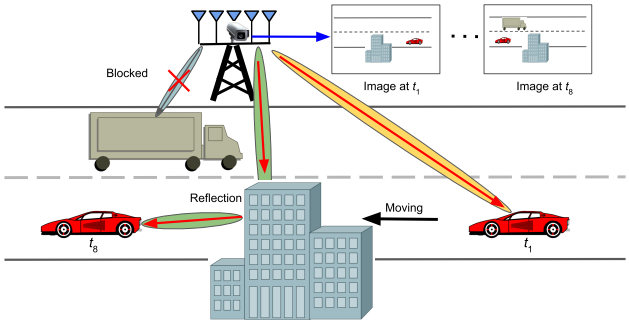

Fig. 2. Scenario of applying DL-based CV to mmWave MIMO beamforming

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 large

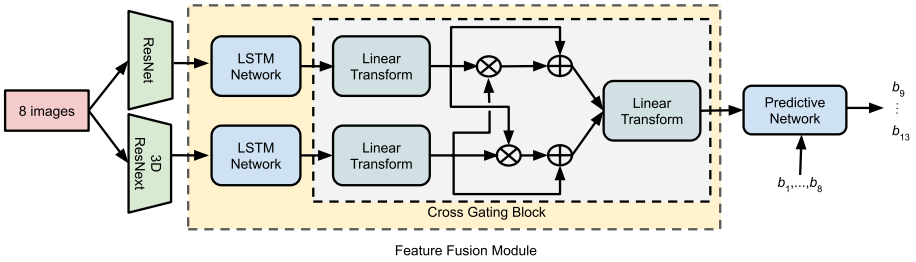

Fig. 3. Architecture of our proposed framework

available bandwidth and ultra-high data-transmitting rate [5]–[7]. Beamforming should be implemented in a large antenna array to achieve the required high power gain and direction. The classic MIMO beamforming 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 a DL-based CV to mmWave MIMO beamforming. A scenario with a BS forming a MIMO beam to serve a target user moving along a street is considered, as shown in Fig. 2. 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 Fig2, 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 red, green, blue (RGB) images of the street view to assist the beamforming process. The problem here is how to utilize the previously-observed eight consecutive beam and their corresponding eight images to predict the future one, three, and five beams. Notably, these beams are represented as beam indices under the same predefined codebook.

B. Framework Architecture and Methods

We propose the DL network framework shown in Fig. 3 composed of ResNet [2], 3D ResNext [8], a feature-fusion module (FFM) [9], and 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 presents the structure of a ResNext block [10], an improved version of a 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, they can focus on more than one specific feature representations.

In 3D ResNext, a similar structure can be observed but with 3D convolutional layers instead of two-dimensional (2D)

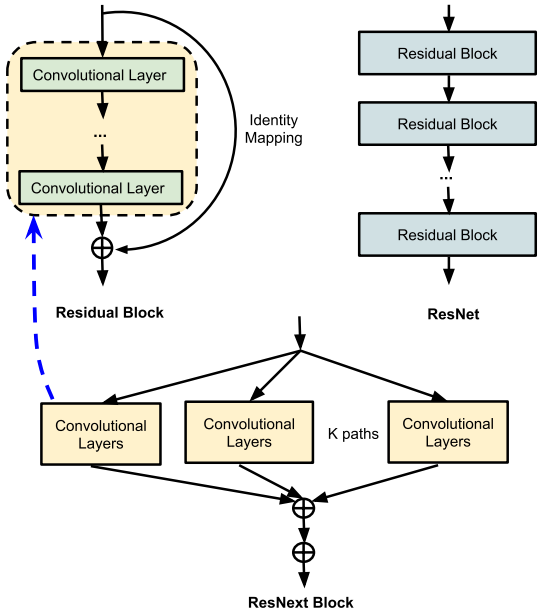

Fig. 4. Structure of residual block, ResNet, and ResNext block

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 becomes extremely long, and many computational resources are occupied due to the many 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 [11]. These features are then fed to a DL network as inputs.

2) Long Short-Term Memory (LSTM) Network:

3) Method with 1D LSTM Network: The LSTM [12] 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), previous long-term memory (hidden state), and previous short-term memory (cell 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. It outputs new long-term memory and short-term memory. The latter is also regarded as a prediction.

When an LSTM cell is recursively utilized in a 1D array

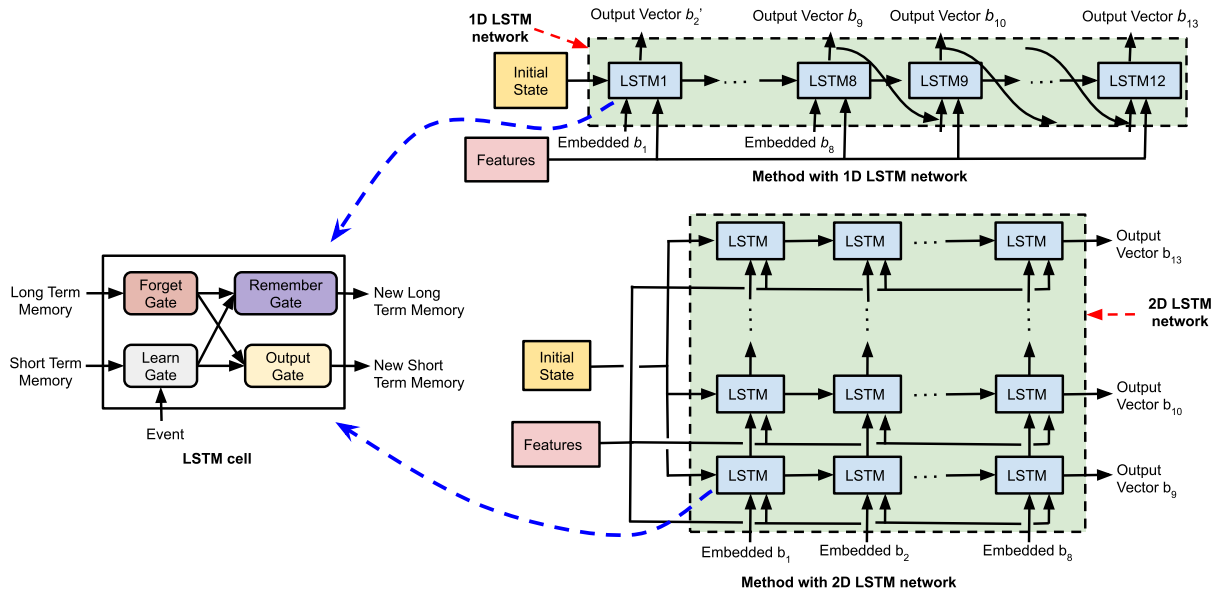

Fig. 5. Structure of LSTM cell, method with 1D LSTM network, and method with 2D LSTM network

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. Using this network, one can yield as many predictions as recursive times.

As shown in Fig. 5, a 2D LSTM network can be realized when the LSTM cell is recursively in a 2D mesh form [13]. 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.

4) Feature-Fusion Module (FFM): Fig. 3 shows the structure of the FFM. It comprises two LSTM networks and a cross-gating block, the former for aggregating these features. High-level features can be obtained through these two networks. The latter 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) to extract visual features. These features are then merged through FFM and sent to the predictive network. Section III-B2 introduces three forms of the LSTM network based on which three methods of designing DL networks are proposed and explained below.

When the predictive network is a 1D LSTM network Fig. 3, 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'. The input and output of each LSTM cell are embedded vectors and output vectors, respectively. Embedding is mapping a constant (beam index) to a vector and can represent the relation between constants well.

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;

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

Step 3: The output 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:

Substep 4.1: The embedded vectors of the first eight beam indices go through the first to eighth 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.

1) 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

TABLE I
PERFORMANCE OF EXPONENTIAL DECAY SCORES AND TOP-1 ACCURACY

	Exponential Decay Score			Top-1 Accuracy		
	1 future beam	3 future beams	5 future beams	1 future beam	3 future beams	5 future beams
Method with 1D LSTM Network	0.9238	0.8206	0.7356	0.9170	0.7719	0.6448
Method with Modified 1D LSTM Network	0.8974	0.7137	0.6129	0.8887	0.6260	0.4800
Method with 2D LSTM Network	0.8803	0.6857	0.5877	0.8704	0.5893	0.4503
Baseline Method in [6]	0.86	0.68	0.60	0.85	0.60	0.50

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.

2) 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 embedded vectors of the first eight beam indices into the LSTM network and 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.

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, and the last five pairs are ground truth label pairs; that is, they have the future beams of the same user and the corresponding images. In this experiment, the first eight pairs serve as the inputs of the designed DL network to generate the predicted future five beam indices to compare with the last five given beam indices.

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 8 epochs. The batch size is set as 256. The cross-entropy loss is utilized for the loss function.

3) Performance: Following the evaluation in [6], the performances of our proposed methods are evaluated on the validation set with the same metrics, which are the exponential decay score and top-1 accuracy. Equations (6) and (7) in [6] explain their detailed expression. Table I lists our results, in which the baseline method in [6] is considered for comparison purposes. In the baseline method, the authors simply leveraged the beam-index data and ignored image data.

From the exponential decay scores, we can see that our proposed 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'.

For the top-1 accuracy, the designed method with the 1D LSTM network also outperforms the baselined method. 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'.

In summary, among the three proposed methods, the method with the 1D LSTM network shows the best beam prediction for the target mobile scenarios.

IV. CHALLENGES AND OPEN PROBLEMS

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

A. Building Datasets

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. Therefore, 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 better to obtain the three sets. These data should be well-organized and easily manipulated, which is, normally, the hardest work in DL 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 tackles optimization problems [14]. GCNs address network-related issues [15], 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 is a kind of RNN that can handle unordered sequences of data. It can replace the 2D LSTM network. 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 [9], [11].

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

The previous sections explained the problem of beam and blockage prediction in a mmWave MIMO communication system. 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: 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. For example, the motion information of the users at the edge of the coverage area can be utilized to forecast and judge whether/when a terminal goes out or comes into its serving area, and then accurate 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.

(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.

(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 and number of terminals.

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 showed 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*. MIT press, 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Pro. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, Jun. 27-30, 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 *2020 10th Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence)*. Noida, Uttar Pradesh, India: IEEE, Jan. 29-31, 2020, pp. 294–299.

[4] W. Xu, F. Gao, S. Jin, and A. Alkhateeb, "3d scene based beam selection for mmwave communications," *arXiv preprint arXiv:1911.08409*, 2019.

[5] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, "Viwi: A deep learning dataset framework for vision-aided wireless communications," *arXiv preprint arXiv:1911.06257*, 2019.

[6] M. Alrabeiah, J. Booth, A. Hredzak, and A. Alkhateeb, "Viwi vision-aided mmwave beam tracking: Dataset, task, and baseline solutions," *arXiv preprint arXiv:2002.02445*, 2020.

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

[8] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, Jun. 18-23, 2018, pp. 6546–6555.

[9] 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. of the IEEE Int. Conf. on Computer Vision*, Seoul, South Korea, Oct. 27-Nov. 2, 2019, pp. 2641–2650.

[10] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Venice, Italy, Oct. 22-29, 2017, pp. 1492–1500.

[11] J. S. Park, M. Rohrbach, T. Darrell, and A. Rohrbach, "Adversarial inference for multi-sentence video description," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, Jun. 16-20, 2019, pp. 6598–6608.

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

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

[14] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[15] 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. of the 2019 ACM Symp. on SDN Research*, San Jose, CA, USA, Apr. 3-4, 2019, pp. 140–151.

VI. BIOGRAPHIES

Yu Tian received his B.Sc in Communication Engineering from the Harbin Institute of Technology in 2013 and his M.Sc degree from King Abdullah University of Science and Technology (KAUST) in 2019. Now he is a Ph.D. student in KAUST. His current research interests include deep learning, performance analysis of wireless communication systems.

Gaofeng Pan (M'12, SM'19) received his Ph.D. degree in Communication and Information Systems from Southwest Jiaotong University, Chengdu, China, in 2011. Since Aug. 2019, he has been a Visiting Researcher with the Communication Theory Lab, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. He has also been with School of Information and Electronics, Beijing Institute of Technology, P. R. China, as a professor, from Apr. 2020. His research interest spans special topics in communications theory, signal processing, and protocol design.

Mohamed-Slim Alouini [S'94, M'98, SM'03, F'09] received his Ph.D. degree in Electrical Engineering (EE) from the California Institute of Technology (Caltech), Pasadena, in 1998. He served as a faculty member at the University of Minnesota, Minneapolis, and then at the Texas A&M University at Qatar, Education City, Doha, before joining KAUST as a Professor of EE in 2009. His current research interests include modeling, design, and performance analysis of wireless communication systems.