IEEE ComSoc IEEE Communications Society | IEEE COMPUTER SOCIETY | IEEE Signal Processing Society | VTS Connecting the Mobile World | IEEE Transactions on Machine Learning in Communications and Networking

Received 23 December 2022; revised 19 April 2023 and 7 August 2023; accepted 15 October 2023.
Date of publication 20 October 2023; date of current version 31 October 2023.

The associate editor coordinating the review of this article and approving it for publication was K. Seddik.

Digital Object Identifier 10.1109/TMLCN.2023.3326409

Computer Vision Aided Beamforming Fused With Limited Feedback

T. XIANG, J. GU, Y. WANG, Y. GAO, AND X. ZHANG (Member, IEEE)

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

CORRESPONDING AUTHOR: X. ZHANG (zhangxin@bupt.edu.cn)

This work was supported in part by the 5G Evolution Wireless Air Interface Intelligent Research and Development and Verification Public Platform Project under Grant 2022-229-220.

ABSTRACT The growing antenna array scale, the uncorrelated fadings between downlink and uplink of frequency division duplex (FDD) or analog beamforming design increases the difficulty of channel sounding or estimation. Non-wireless channel detection or beam weight prediction method is a promising solution to help obtain timely and accurate wireless channel state. Beamforming can be enhanced by the powerful sensing capability of cameras, for which this paper proposes a straightforward beam weight prediction method by implementing convolutional neural network (CNN) on images from cameras and a loss function based on chordal distance for this paper's task. Then a fusion method of visual detection and wireless sounding is developed to further improve spectral efficiency. This fusion method also utilizes a codeword rotation mechanism with Householder transform to save the notification overhead of visual detection results. A testbed has been built to verify the proposed approach with field measurement data. The proposed straightforward method is able to reach high spectral efficiency performance, and the fusion method could outperform exclusive visual detection or wireless sounding with appropriate hierarchical codebook.

INDEX TERMS Beamforming, computer vision, hierarchical codebook.

I. INTRODUCTION

With the scale of antenna array growing larger, channel estimation will cost increasing computational and wireless spectrum resources [1], which calls for introduction of non-wireless detection methods, such as visual detection [2]. Moreover, beam weight prediction performed by visual detection could also promise a way to overcome the inherent defect for frequency division duplex (FDD) mode [3] or analog beamforming to obtain timely and accurate wireless channel status [4].

To address the problems above, computer vision (CV) aided beamforming has become a feasible solution. In our previous work, CV was utilized to select mmWave beams in LOS scenario with no or limited wireless overhead, for mmWave signals have good directivity and similar propagation characteristic as visible light [5]. In our work, visual detection also can save lots of overhead in various scenes, such as electromagnetic exposure control [6] for high-gain arrays and over-the-air phase calibration [7] for phased arrays. ViWi (Vision-Wierless) [8] is a data-generating framework that does not only provide wireless data but combines it with visual data taken from the same scenes, which is enabled by 3D modeling and ray-tracing simulators that generate high-fidelity synthetic vision and wireless data. Based on this simulation-generated dataset, researchers propose a vision-based beam blockage prediction method, which utilizes convolutional neural network (CNN) to predict whether a communication target is blocked [9] as well as whether it should be switched to another high frequency base station [10] with captured RGB image. Furthermore, beam selection could also be implemented by vision detection with the help of CNN, which achieves high accuracy of mmWave beam and blockage prediction [11]. The vision aided beam selection has also been realized under multi-user scenario, with user selection network as the first step and beam selection network as the second step [12]. Apart from images captured by camera, LiDAR (Light Detection And Ranging) is also powerful in guiding beam selection. In [13], a mmWave beamforming communication system fused with LiDAR and camera has been built, where blockage is detected via vision or LiDAR in order to instruct frequent handoff. In [14], a LiDAR based fast mmWave beam search approach has been proposed, which

reduces beam search time and overhead in mmWave vehicle-to-infrastructure scenario. However, current researches focus on the combination of vision detection and the wireless signal directivity, especially when mmWave or THz band is applied. This might become degraded or even inapplicable in terms of prediction performance when the wireless signal loses good directivity, or in other words, the sparsity characteristic [15], i.e., as in the case of sub-6GHz band. In addition, these studies all regard beam prediction as a classification prediction task, considering that codebook-based beam selection is the most commonly used method in mmWave communications, but its prediction performance is also limited by the discrete beam selection mechanism.

In fact, applying universal approximation theory [16] on the prediction of wireless channel might be a powerful solution of building the bridge between vision detection and wireless detection. Utilizing the assumption that wireless channel can be mapped from geometric position distribution, wireless channel can be predicted by using geometry-related measurements via deep neural network. Since the mapping function exists at different frequencies, the wireless channels at different frequency bands could also establish connections with each other, which enables mutual prediction between uplink and downlink channels in FDD system [17]. Similarly, researchers in [18], [19], and [20] leverage the spatial correlation between the sub-6 GHz and mmWave channels to help reduce the high mmWave beam training overhead and maintain reliable links with blockages. Object detection task in CV is regarded as a typical positioning method, so it can be a positioning technique for channel awareness purposes. In [21], the wireless channel covariance matrix is proved to be predictable based on scene image with user in it, and the prediction task is implemented by CNN with simulation generated dataset. The theoretical foundation of [9] and [11] is also based on the mapping from user's location to optimal beam index. Although CV aided channel covariance matrix and optimal beam prediction can bring indeed performance improvements in multi-antenna communication, whether the wireless channel itself or optimal beamforming weights can be predicted with CV detection still needs theoretical analysis and experimental verification.

Furthermore, other studies have not considered fusion methods of vision and wireless detection, relying exclusively on ML methods, which is not likely to take advantage of wireless feedback once it were available.

This paper proposes a CV aided straightforward beam weights prediction method, regarding beam prediction as a regression task for broader frequency band generality. And a hierarchical limited wireless feedback mechanism fused with CV detection is introduced to enhance the performance. The contributions of this paper are as follows:

- This paper reveals the mapping function from coordinates of nodes in visual detection to wireless channel/optimal beam weights. And we prove that under the assumption that the mapping function from geometric position to wireless channel exists, the prediction of optimal beam weights can reach any arbitrary precision with the knowledge of user equipment's (UE) image coordinates, and also enables approaching the maximal achievable transmission rate with perfect channel information.
- This paper proposes to use convolutional neural network for constructing the bridge from captured RGB images to optimal beam weights rather than beam index with free pilot or feedback overhead. Visual Geometry Group (VGG) network, one of typical forms of implementation for CNN, is applied to accomplish this task. A loss function based on chordal distance is proposed and performs better than MSE loss function.
- A method combining CV and limited wireless feedback is proposed to enhance the beam weight predicting precision. CV prediction promises an overhead-free but low-resolution weight prediction, meanwhile the second step of hierarchical wireless feedback would raise the detection resolution thus improving the spectral efficiency. Then a codeword rotation mechanism is proposed, which applies Householder transform to save the overhead for CV prediction result notification. The performance under different SNR, dataset and wireless feedback overhead is discussed, indicating that the proposed fusion method outperforms individual CV or wireless detection under certain conditions.
- A testbed of vision aided beamforming has been built to collect image-wireless channel related dataset in an indoor scenario at sub-6GHz frequency band. And the proposed approach is verified by the field measurement data. Beyond that, two sets of environment data have been collected, which are used for verifying the proposed method's generality for different environments with the help of transfer learning.

The paper is organized as follows. Section II describes the problem and system model for CV-aided beamforming system. Section III introduces the feasibility of vision-based beam weight prediction and proposes a straightforward prediction method implemented by CNN. Section IV introduces a hierarchical limited feedback method fused with visual detection. Section V presents our testbed for wireless and vision fusion experiment, and then shows the experiment results and analysis. And finally, section VII draws the conclusion.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In a MISO system, the received signal can be expressed as:

$$y = \boldsymbol{h}^H \boldsymbol{w} x + n \tag{1}$$

wherein $\boldsymbol{h}$ is the wireless channel, $\boldsymbol{w}$ denotes the beam weight vector, x is the transmitted signal and n is the noise. With SNR ρ, the achievable data rate with analog beamforming can be

expressed by [17]:

$$R(\boldsymbol{h}, \boldsymbol{w}) = \log_2\left(1 + \rho\left|\boldsymbol{h}^H \boldsymbol{w}\right|^2\right) \tag{2}$$

To maximize the achievable rate, the beam weight is represented as:

$$\begin{aligned} \boldsymbol{w}^* = &\underset{\boldsymbol{w}\in C^{M\times 1}}{\operatorname{argmax}} R(\boldsymbol{h}, \boldsymbol{w}) \\ &\text{s.t. } \boldsymbol{w}^H \boldsymbol{w} = 1 \end{aligned} \tag{3}$$

However, the acquisition of wireless channel is usually difficult especially when the number of elements in antenna array grows larger. Due to the large array scale, the training overhead could be costly. Moreover, FDD transmission design or analog beamforming would also limit the accuracy of channel acquisition because of the limitation of sounding and feedback mechanism. TDD systems, on the other hand, are also subject to reciprocity loss, as the different transceiver circuits in TDD systems can lead to asymmetric links, especially if there are calibration errors in large-scale arrays [22], besides the loss from time-varying effect of the channel. So in this paper, we are going to look into this problem from another perspective [5], [11]. Wireless channel could be predicted by RGB images captured by cameras [11], [13], [19]. But different from [11], [13], and [19], this paper intents to directly predict optimal beam weight rather than select if from predefined codebook, thereby enabling the visual detection to cooperate with beam training process. As for the difference between the deployment in FDD and TDD systems, the former could directly use the proposed method due to its inherent design, while the latter should choose whether or not to use feedback for channel acquisition and the corresponding ML method based on the degree of circuit error.

This paper intends to minimize the difference between achievable data rate with optimal beam weight and the data rate with vision-based predicted weight. The optimization problem is represented as:

$$\begin{aligned} &\min_{\hat{\boldsymbol{w}}\in C^{M\times 1}} R\left(\boldsymbol{h}, \boldsymbol{w}^*\right) - R\left(\boldsymbol{h}, \hat{\boldsymbol{w}}\right) \\ &\text{s.t.} \quad \hat{\boldsymbol{w}}^{\mathrm{H}} \hat{\boldsymbol{w}} = 1 \end{aligned} \tag{4}$$

wherein $\hat{\boldsymbol{w}}$ is the vision-based predicted weight.

Optimization problem (4) has been used for finding efficient and accurate wireless feedback approaches to maximize the communication spectral efficiency in typical wireless communication systems. In this paper, we propose to utilize CNN to predict beam weight, and further propose a hierarchical vision-wireless fusion method, which continues to follow (4).

III. VISION-BASED BEAM WEIGHT PREDICTION

In this section, we first state the feasibility of the vision-based beam weight prediction, and then we introduce our straightforward prediction method from captured images to beam weight, which could save wireless feedback and pilot overhead.

A. IMAGE COORDINATE TO BEAM WEIGHT MAPPING

Let $\mathbf{c}_{w,3\mathrm{D}} = [x_w, y_w, z_w]^{\mathrm{T}} \in R^{3\times 1}$ be the real world coordinate of UE, $\mathbf{c}_i = [x_i, y_i]^{\mathrm{T}} \in R^{2\times 1}$ be the coordinate of UE in image coordinate system, s be the distance between camera and object, f be the focal length of the optical system, $\mathbf{R}$ be the rotation matrix and $\mathbf{T}$ be the translation matrix [23]:

$$\mathbf{c}_{w,3\mathrm{D}} = s\mathbf{R}\begin{bmatrix} x_i/f \\ y_i/f \\ 1 \end{bmatrix} + \mathbf{T} \tag{5}$$

$$\begin{aligned} \mathbf{R} = &\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix}\begin{bmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{bmatrix} \\ &\times \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{6}$$

It can be derived that when the UE's height is a constant h_0, the 2D world coordinate of UE at height h_0 is a function of $\mathbf{c}_i$. It should be noted that the world coordinate can be uniquely determined by monocular vision if h_0 is constrained. And this paper focuses on monocular vision. Even if the UE height is not fixed, binocular vision is also capable of 3D positioning [24], [25]. So there exists a function mapping image coordinate to world coordinate, represented as:

$$\Theta : \{\mathbf{c}_i\} \rightarrow \left\{\mathbf{c}_{w,2D}\right\} \tag{7}$$

Here $\left\{\mathrm{c}_{w,2\mathrm{D}}\right\} = \{(x_w, y_w, z_w) \,|x_w, y_w \in R, z_w = h_0\} \subset \left\{\mathrm{c}_{w,3\mathrm{D}}\right\}$. Without losing generality, the origin of real world coordinate system is set as the bottom of the base station at ground level, and the origin of the image coordinate system is set as the top left corner of the image. Since the coordinate transformation is performed by affine transformation, the translation of origin does not affect the approach.

From the perspective of wireless ray propagation, after the wireless signal is transmitted from the transmitter, it arrives at the receiver after line-of-sight, reflection, diffraction and scattering transmission. Each transmission path is superimposed to form the amplitude and phase of the received signal. These propagation paths have different intensity attenuations and phase variations due to varied modes of propagation as well as delay. The expression for the superposition of these paths to form the channel of MISO system is as follows:

$$\boldsymbol{h} = \sum\nolimits_{l=1}^{L} \alpha_l e^{j\varphi_l} \boldsymbol{a}\left(\theta_l^a, \theta_l^e\right) \tag{8}$$

where α_{l}, φ_{l}, $\theta_{\mathrm{l}}^{\mathrm{a}}$ and $\theta_{\mathrm{l}}^{\mathrm{e}}$ denote the amplitude attenuation, phase changing, azimuth of departure and elevation of departure for the l-th path, and a $\left(\theta_{\mathrm{l}}^{\mathrm{a}}, \theta_{\mathrm{l}}^{\mathrm{e}}\right)$ is the steering vector of the antenna array.

The propagation of all paths can be seen as a result of the interaction between the wireless signal and the environment. When the relative position changes, the presence or absence of paths, their propagation mode, angle and propagation distance will change. This indicates that the distribution, the amplitude and phase of each path are closely related to the relative position of the transmitter, the receiver and the

environment. Thus in a given environment, when the base station's location is fixed, the location of the mobile terminal determines the wireless channel between them. Based on that, this paper assumes the existence of mapping from terminal location to wireless channel, which has also been corroborated by [18], [19], and [20]:

$$\Phi_{2D} : \left\{\mathbf{c}_{w,2D}\right\} \to \{\boldsymbol{h}\} \tag{9}$$

Assumption 1: There exists a unique optimal solution for problem (4). Define the mapping function (assumed to be continuous) from channel to the optimal weight as:

$$\Xi : \{\boldsymbol{h}\} \to \left\{\boldsymbol{w}^*\right\} \tag{10}$$

Assumption 1 is needed to illustrate the feasibility of beam weight prediction using the universal approximation theorem, as discussed shortly below. Although the problem (4) has no closed form solution, assumption 1 can be approximately obtained under flat fading channel assumption, i.e., the optimal beam for a MISO narrowband system is $\boldsymbol{w}^* = \boldsymbol{h}/\|\boldsymbol{h}\|$. Therefore, function Ξ exists and is continuous.

Proposition 1: Under assumption 1, there exists an image coordinate to optimal beam weight function:

$$\Psi : \{\mathbf{c}_i\} \to \left\{\boldsymbol{w}^*\right\} \tag{11}$$

Proof: the proof of existence of Ψ follows from the existence of image to position mapping function $\Theta^{-1}(\cdot)$, position to channel mapping function $\Phi_{2D}(\cdot)$ and channel to the optimal beam weight mapping function $\Xi(\cdot)$ [17].

Based on proposition 1, problem (4) can be rewritten as

$$\min R\left(\boldsymbol{h}, \boldsymbol{w}^*\right) - R\left(\boldsymbol{h}, \hat{\Psi}\left(\mathbf{c}_i\right)\right) \tag{12}$$

wherein $\hat{\Psi}(\cdot)$ represents an artificial approximation function of $\Psi(\cdot)$, which could be a neural network.

Proposition 2: For $\forall\varepsilon > 0$,there exists a neuron number N that satisfies:

$$\sup_{\mathbf{c}_i} \left\| \prod\nolimits_N \left(\mathbf{c}_i, \Omega\right) - \Psi\left(\mathbf{c}_i\right) \right\| < \varepsilon \tag{13}$$

where $\prod_N(\mathbf{c}_i, \Omega)$ denotes the outputs of a neural network that consists of one hidden layer with N neurons and Ω is the network parameters.

Proof: The image coordinate $\mathbf{c}_i$ is bounded and closed, and $\{\mathbf{c}_i\}$ is a compact set. The bounds on $\mathrm{c_i}$ is determined by the actual image boundaries, e.g. for a WxH image, $\mathrm{c_i} = \left\{(\mathrm{x_i}, \mathrm{y_i})^{\mathrm{T}} \,|\mathrm{x_i} \in [0, \mathrm{W}], \mathrm{y_i} \in [0, \mathrm{H}]\right\}$. (ii) $\Psi(\mathbf{c}_i)$ is a continuous mapping function. Therefore, $\prod_N(\cdot, \Omega)$ can be proved by universal approximation theorem [16] to be able to approximate $\Psi(\mathbf{c}_i)$ with arbitrary accuracy.

Corollary 1: For $\forall\varepsilon > 0$,there exists a neuron number N that satisfies:

$$\sup_{\mathbf{c}_i} R\left(\boldsymbol{h}, \boldsymbol{w}^*\right) - R\left(\boldsymbol{h}, \prod\nolimits_N \left(\mathbf{c}_i, \Omega\right)\right) < \varepsilon \tag{14}$$

Proof: Considering that $\log_2\left(1 + \rho\left|\boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{w}\right|^2\right)$ is continuous for $\boldsymbol{w}$, proposition 2 reveals that $\sup_{\mathbf{c}_i} \left\|\boldsymbol{w} - \boldsymbol{w}^*\right\| < \varepsilon$, so corollary 1 is obtained.

Proposition 2 and corollary 1 provides the theoretical basis for this paper's task, which indicates that beam weight prediction could be achieved with monocular vision detection by neural network with arbitrary precision. From a practical point of view, however, the image coordinate also needs to be extracted from captured images. So we adopt a straightforward design, realizing image coordinate extraction and beam weight mapping in one convolutional neural network, which will be described in the next subsection.

B. STRAIGHTFORWARD PREDICTION METHOD WITH CNN

We assume that real-time images captured by a monocular vision camera is fed to a CNN, and then the optimal beam weights related to the captured UE is directly predicted by the CNN. Since the typical CNN network is suitable for object detection or positioning task, the image coordinate could be obtained easily [26]. The underlying principle of this paper is to combine the positioning task and beam weight predicting task by regarding the positioning task as an implicit form, so that the latter part of the network could represent the beam weight predicting task.

While the typical multi-layer perceptron (MLP) can be a basic approximator design, CNN is more suitable for tasks where 2D images are input, because the convolutional structure is designed specifically for extracting 2D image features. In addition, the fully-connected (FC) layers in the CNN design are expected to perform the mapping of image features to beam weights, just as MLP does.

The network design in this paper follows the principle of VGG network [26], including several convolution-pooling layers, followed by FC layers. The 2D convolutional layer can be expressed as:

$$z_{u,v}^{(l)} = g\left(\sum\nolimits_{i=1}^{W}\sum\nolimits_{j=1}^{H} x_{p,i+u,j+v}^{(l)} \cdot k_{i,j}^{(l)} + b^{(l)}\right) \tag{15}$$

wherein $g(\cdot)$ is the activation function and the $x_{p,i+u,j+v}^{(l)}$ is the padded image input, which is calculated as:

$$x_{p,i,j}^{(l)} = \begin{cases} x_{i,j}^{(l)}, & i < M \\ 0, & otherwise \end{cases} \tag{16}$$

And the fully-connected layer is calculated as:

$$\boldsymbol{x}^{(l)} = g\left(\boldsymbol{w}_{FC}^{(l)} \cdot \boldsymbol{x}^{(l-1)} + \boldsymbol{b}^{(l)}\right) \tag{17}$$

The basic network structure follows the VGG network, as shown in figure 1, which will be covered in detailed in section V. Considering that the predicted weight output has a normalized power constraint, the activation function of the output layer is expressed as:

$$\boldsymbol{x}_{output} = \frac{\boldsymbol{x}_{\text{input}}}{\left\|\boldsymbol{x}_{\text{input}}\right\|_2} \tag{18}$$

Except that, Leaky Relu is adopted as activation function in other layers, including convolutional layers and FC layers.

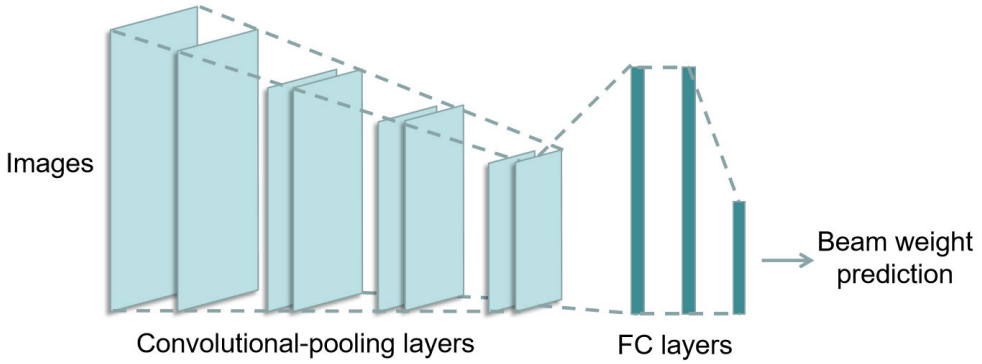


FIGURE 1. **CNN structure.**

We then introduce the following steps in our method: dataset gathering, network training and performance evaluation.

1) DATASET GATHERING

This step gathers dataset for neural network training. The dataset consists of image inputs and beam weight labels which is presented as $D = \left\{\left(\boldsymbol{Y}^{(1)}, \boldsymbol{w}^{(1)}\right), \left(\boldsymbol{Y}^{(2)}, \boldsymbol{w}^{(2)}\right), \ldots, \left(\boldsymbol{Y}^{(K)}, \boldsymbol{w}^{(K)}\right)\right\}$, where K is the dataset cardinal. Part of the dataset is shared in [27], which is collected by our testbed described in section V.

In order to derive the optimal beam weight matched with a fixed location, we apply uplink feedback method. Due to channel reciprocity, uplink signal transmitted by UE can be used to estimate wireless channel. Then $\boldsymbol{w}^{(K)}$ is derived from wireless channel. Meanwhile, an image $\boldsymbol{Y}^{(K)}$ containing the UE is captured by the camera.

It should be noted that the channel reciprocity is considered in dataset gathering step to study the performance limit of the proposed method. Furthermore, the labels could also be designed as quantized values rather than accurate ones, since data set might be gathered exclusively in wireless systems with FDD or analog beamforming design. In this case, the network becomes similar to classifiers in [11], [13], and [19], but has numerical outputs. Although channel reciprocity might be lost due to FDD mode or disabled digital channel estimation due to analog beamforming design, as long as there is high precision channel feedback collected in advance or gathered by part of high performance nodes, the proposed method is expected to outperform these classifiers.

For dataset preprocessing, all the samples are shuffled to overcome the regularity brought by sample collection. Meanwhile the optimal weight with measured normalized channel is derived.

2) NETWORK TRAINING

After dataset has been gathered, we train the neural network to minimize the loss function. Since the task in this paper is a regression task, the loss function can be MSE, expressed as:

$$J(\Omega) = \frac{1}{K}\sum_{k=1}^{K}\left\|\hat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)}\right\|_2^2 \tag{19}$$

Or the loss function can also be defined by chordal distance [28]:

$$J(\Omega) = \frac{1}{K}\sum_{k=1}^{K}\sqrt{1 - \left|\left[\hat{\boldsymbol{w}}^{(k)}\right]^{\mathrm{H}} \cdot \boldsymbol{w}^{(k)}\right|^2} \tag{20}$$

In this paper's task, although the optimal beam weight $\boldsymbol{w}^*$ in assumption 1 is unique, a random phase rotation $e^{j\theta}$ of beam weight will not change the achievable data rate nor the chordal distance. So the chordal distance loss function encourages the network to learn rotated weights rather than the unique optimal $\boldsymbol{w}^*$ which is represented by MSE loss function.

It is important to note that the rotation of the complex weights in the complex plane does have no effect on the chordal distance, as stated in [29], which also indicates that the principle of codeword optimization is to maximize the minimum chordal distance between the rotations of a codeword in the complex plane and other codewords. The prediction task in this paper is similar that it is sufficient to predict the arbitrary rotations in the complex plane of a co-phased weight.

The MSE loss function predicts the single co-phased weight and is not inclusive of rotations in the complex plane, whereas the chordal distance loss function treats all rotations in the complex plane equally and will encourage the neural network to learn weights that are closer in chordal distance, even though the Euclidean distance may be larger.

3) PERFORMANCE EVALUATION

When CNN training has been conducted, the system is deployed with beam weight vector predicted by CNN and camera rather than estimated by pilot or obtained by codebook-based search. This method could promise to save the time-frequency wireless overhead for channel estimation or beam training.

Chordal distance is a direct metric that validates how accurate a CNN predicts. Since chordal distance stands for

the correction extent between predicted weight and optimal weight, so this metric directly reflects the prediction precision.

To validate the beam weight prediction precision, spectral efficiency is chosen as the key indicator. If the transmitted power is fixed, a good transmission weight would lead to high spectral efficiency. So this metric implements the CNN prediction performance to the indicator of communication purpose.

IV. HIERARCHICAL CODEBOOK WITH CV PREDICTION DESIGN

Although CV may provide high prediction precision, a large amount of CNN layers and neurons are needed for precision enhancement. Therefore, high precoding gain requires large scale CNN, which costs lots of calculation resources and processing time. However, wireless overhead could provide high detection precision, if high-resolution codebook can be constructed. So we consider combining CNN with wireless sounding and feedback to obtain higher precoding gain meanwhile cutting down wireless feedback overhead. We expect that introducing a few wireless feedback overhead to the CV method could promote the performance.

As mentioned earlier, CV may not be capable of predicting high-resolution quantized beamforming weights, so wireless feedback should be considered to improve beam weight quantization resolution. We now propose a hierarchical codebook mechanism aided by CV. The hierarchical design is common in pure wireless sounding to raise the detection resolution [30], [31], which generally includes two tiers of wireless sounding. In this method, to obtain the optimal beam weight, two steps are needed. First, apply the CNN to predict an initial weight, for which the precision may be constrained by CNN scale. Second, inform UE about the initial weight with certain quantization. The resolution is determined by the codebook quantization bit number. Finally, the UE estimates the wireless channel, and feeds back a fine-resolution beam weight. Based on the low-resolution beam weight predicted by CV, UE feedback is supposed to further improve beam weight resolution. However, the CV-predicted beam weight may be any value, rather than codewords in the parent codebook. So how to match CV prediction and hierarchical codebook is the key to combine CV prediction and wireless feedback method.

The parent codeword matching mechanism is to select a codeword that has the closest distance to the CV predicted weight in the parent codebook. However, this may cause problems in overhead cost. The base station (BS) needs to notify the UE of the predicted parent codeword, which needs certain wireless overhead. This paper proposes a codebook rotation mechanism to overcome this problem.

The codebook rotation mechanism is used to avoid the cost of notifying CV of predicted parent codeword. Knowing the predicted codeword based on CV, when performing wireless detection in the second step, the transmission beamforming controller first adjusts the wireless channel according to the expected default codeword. This process is implemented via Householder transform. As a typical linear transform, Householder transform could be used to construct a unique codebook with a common base codebook [32], [33]. First, the BS and UE should confirm a fixed codeword, for instance, the first codeword in a parent codebook. Then calculate the Householder matrix to transform CV predicted weight to the chosen codeword:

$$\mathcal{H} = \mathbf{I} - 2\frac{\left(\hat{\boldsymbol{w}} - \boldsymbol{w}_1^{\mathrm{cw}}\right)\left(\hat{\boldsymbol{w}} - \boldsymbol{w}_1^{\mathrm{cw}}\right)^H}{\left\|\hat{\boldsymbol{w}} - \boldsymbol{w}_1^{\mathrm{cw}}\right\|^2} \tag{21}$$

Next, the child codeword for wireless sounding is left multiplied by Householder matrix, so that the child codeword is transformed into:

$$\boldsymbol{w}^{\mathcal{H},\mathrm{cw}} = \mathcal{H}\boldsymbol{w}^{\mathrm{cw}} \tag{22}$$

In this way, the child codewords are transformed to be near the predicted weight, as presented in figure 2.

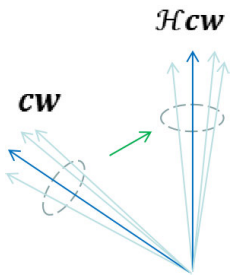


FIGURE 2. **Codebook rotation mechanism.**

The rotation is feasible because at this time, the CV has had a rough knowledge of the channel. In this way, UE only needs to perform accurate wireless detection near a fixed channel, which will only produce the overhead of this part. This paper now analyzes the influence of Householder transform on the child codebook.

Since the Householder matrix is unitary, the Householder transform does not change the coherence-based distance measures. This reveals two advantages of Householder transform. First, the codebook performance remains the same after Householder transform, because the distances between any codeword remain the same. Second, the prediction error of CNN is also maintained in terms of chordal distance, because the distance between real optimal weight and the predicted one is also unchanged. So any codeword in the parent codebook could be chosen as the default codeword without performance degradation.

When it comes to the influence of CV prediction error, it can be considered that wireless channel that is distributed in the vector space is converged near a predicted channel value without wireless overhead. However, CNN itself has prediction error, as well as the sampling equipment is imperfect for measurement, which leads to fluctuation of predicted weight around the optimal one. Therefore, what kind of child codebook should be used needs to be studied, which not only

is supposed to resist the error of the first step measurement, but also assures the detection accuracy.

In this paper, the child codebook is generated by parent codebook. The parent codebook applied is Grassmannian codebook, which uniformly quantizes the codeword space. The Grassmannian codebook is generated by solving the Grassmannian packing problem, which seeks to maximize the minimum distance between codewords [34], [35], [36]. Three generation methods are applied: double centroid (DC), single centroid (SC) as well as half double centroid and half single (HDHS) centroid. To obtain child codewords, the calculation of centroid [37] of two codewords is the key operator. For double centroid method, for instance, choose codeword $\boldsymbol{w}_i^{\mathrm{cw}}$ and $\boldsymbol{w}_j^{\mathrm{cw}}$, the centroid of these codewords is calculated as:

$$\boldsymbol{a}_{i,j} = \mathrm{centroid}\left(\boldsymbol{w}_i^{\mathrm{cw}}, \boldsymbol{w}_j^{\mathrm{cw}}\right) \tag{23}$$

wherein $\boldsymbol{a}_{i,j}$ is the bound of these two codewords, which has equal distance to them. DC codebook intends to generate high resolution codebook, so get centroid again for a child codeword:

$$\boldsymbol{b}_{i,j} = \mathrm{centroid}\left(\boldsymbol{w}_i^{\mathrm{cw}}, \boldsymbol{a}_{i,j}\right) \tag{24}$$

$\boldsymbol{b}_{i,j}$ stands for the centroid of $\boldsymbol{w}_i^{\mathrm{cw}}$ and $\boldsymbol{a}_{i,j}$, which is regarded as a DC child codeword. And all the child codewords can be presented as:

$$\boldsymbol{B}_{DC,i} = \left\{\boldsymbol{b}_{i,j} | j \neq i\right\} \tag{25}$$

As for SC, the child codebook is constructed with the bounds $\boldsymbol{a}_{i,j}$, rather than $\boldsymbol{b}_{i,j}$. The child codebook for $\boldsymbol{w}_i^{\mathrm{cw}}$ is presented as:

$$\boldsymbol{B}_{SC,i} = \left\{\boldsymbol{a}_{i,j} | j \neq i\right\} \tag{26}$$

And for the compromise codebook HDHS, the child codebook of $\boldsymbol{w}_i^{\mathrm{cw}}$ each extracts half codewords from $\boldsymbol{B}_{DC,i}$ and $\boldsymbol{B}_{SC,i}$ respectively. Figure 3 shows the single centroid and double centroid codewords for $\boldsymbol{w}_i^{\mathrm{cw}}$ and $\boldsymbol{w}_j^{\mathrm{cw}}$.

In general, pure wireless hierarchical codebook applies DC method as in this paper. But considering CV prediction error, error-tolerant codebook should be introduced, such as SC and HDHS codebook. Nevertheless wireless approach may also apply overlapping child codebook [38], in order to enhance tolerance of parent codeword selection error caused by low SNR.

It is supposed that SC has the highest tolerance of first step detection error, but has the lowest detection resolution among the three codebooks. It is because with the same amount of quantization bit, the codeword spacing of SC is the largest. On the contrary, DC has the highest resolution but the lowest tolerance. And HDHS seeks a compromise.

V. TESTBED DEPLOYMENT AND EXPERIMENT RESULT ANALYSIS

This section depicts the structure and main components of the testbed, as shown in figure 4. This testbed includes the following components: two USRPs (Universal Software Radio Peripheral) at half duplex mode for wireless channel sounding and data collection, a monocular camera capturing scene images with UE and a dataset gathering server which also controls the RF front-end and the camera. The real world dataset gathering testbed is presented in figure 5.

During the dataset gathering process, the control server controls the RF front-end to obtain amplitude and phase response of wireless channel and meanwhile operates the camera to capture scene image containing UE illustrated as in figure 6. At the same time, the location of the UE is recorded utilizing inertial sensors on the self-controlled

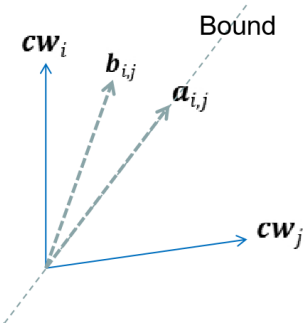


FIGURE 3. The single centroid and double centroid codewords.

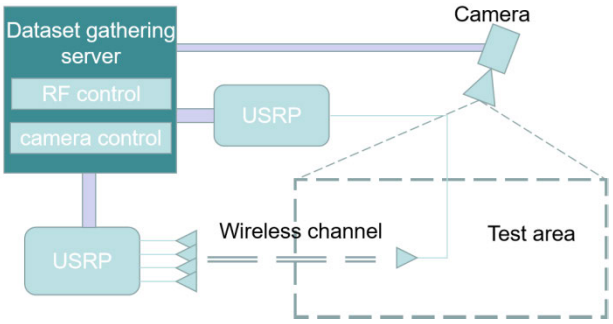


FIGURE 4. The structure and main components of the testbed.

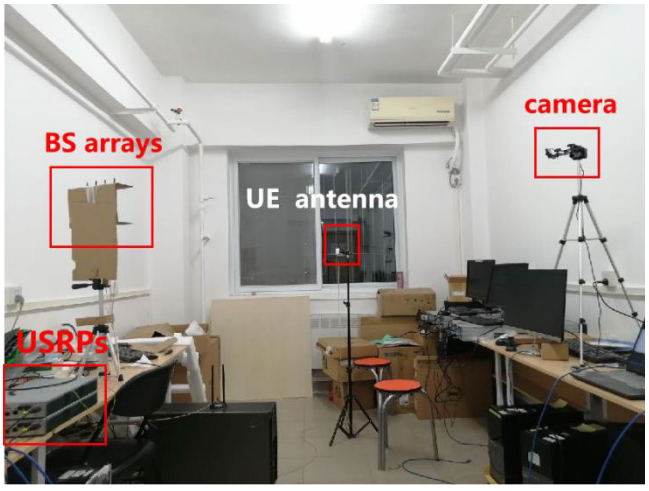


FIGURE 5. The real world testbed.

(a) Scene A image

(b) Scene B image

FIGURE 6. Images captured by camera.

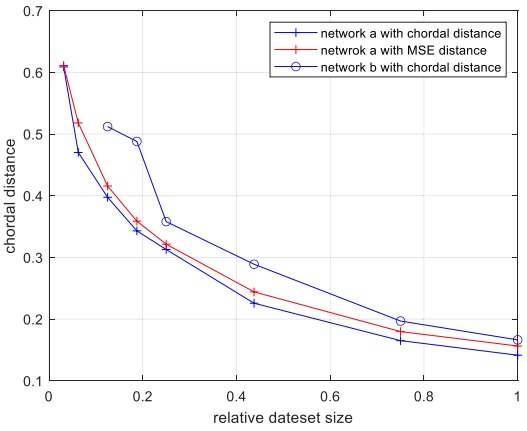


FIGURE 7. Avg chordal distance of visual predicting for different loss functions and networks.

moving platform. In this way, the dataset is generated with samples at random locations.

The CNN structures are illustrated in table 1 and the training hyper-parameters are listed in table 2. "224 × 224, 16" for a convolutional layer means the input size of features are 224 × 224 and the channel number is 16. The size of convolution kernel is 3 × 3. Max pooling is adopted. And FC layers have 1024 or 256 neurons. The activation function for all the layers of CNN adopts Leaky ReLu, except for the output layer which adopts power normalization function.

Loss function based on chordal distance promises a better performance than MSE loss function as presented in figure 7.

TABLE 1. The structure of CNN.

Layers	Network a	Network b
Convolutional	224 x 224, 16	224 x 224, 16
Pooling	2 x 2 Max	2 x 2 Max
Convolutional	112 x 112, 32	112 x 112, 32
Pooling	2 x 2 Max	2 x 2 Max
Convolutional	56 x 56, 64	56 x 56, 64
Convolutional	56 x 56, 64	56 x 56, 64
Pooling	2 x 2 Max	2 x 2 Max
Convolutional	28 x 28, 128	28 x 28, 128
Convolutional	28 x 28, 128	28 x 28, 128
Pooling	2 x 2 Max	2 x 2 Max
Convolutional	14 x 14, 256	14 x 14, 256
Convolutional	14 x 14, 256	14 x 14, 256
Convolutional	14 x 14, 512	14 x 14, 512
Pooling	2 x 2 Max	2 x 2 Max
FC	1024	256
FC	1024	256
FC	1024	256
Output	8	8

TABLE 2. The hyper-parameters for CNN training.

Hyper-parameters	Value
Initial learning rate	0.0001
Batch size	64
Learning rate decay	exponential
Learning rate decay rate	0.965 per epoch
Epoch	30

It is because the MSE loss function looks for certain mapping from captured RGB images to one beam weight, and the task in this paper, however, allows beam weight rotation, leading to non-unique optimal weight. The chordal distance based loss function could encourage the network to predict the rotated weights, which allows for more variation of predicted weights that are valid. So the following analysis is based on the chordal distance loss function. In addition, figure 7 shows that a larger neural network scale does improve the performance of prediction, just as network a outperforms network b. The following research will be based on network a.

Dataset A with 9500 samples without blockage is collected, as shown in figure 6(a), where 8000 samples are for training, 750 for validating and 750 for testing. It is seen in figure 8 that with the relative dataset size increasing, the predicting precision rises, which is represented by the chordal distance between the optimal beam weight label and the predicted one. Large scale network could promise better performance, which is verified from the better performance of network a over b. In the following verification, we use network a for further research.

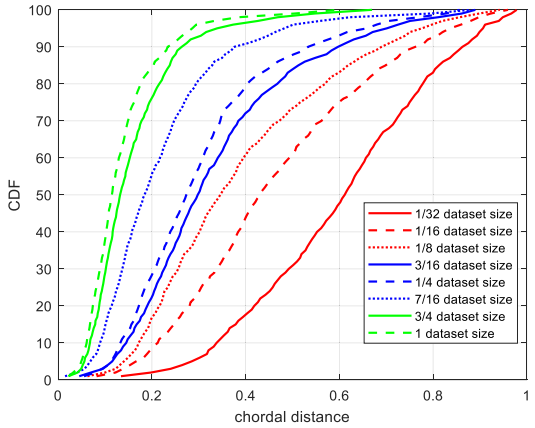


FIGURE 8. Chordal distance CDF of visual predicting with different dataset size.

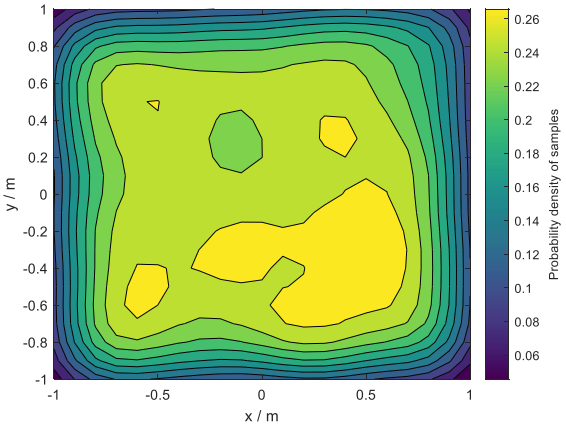


FIGURE 9. Sample density.

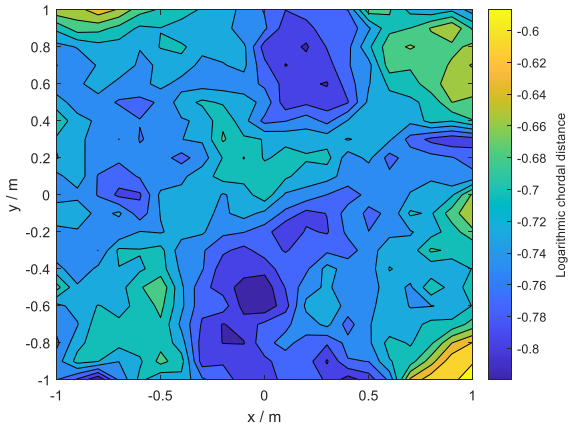


FIGURE 10. Geo-related logarithmic chordal distance.

It can also be revealed that the sample density of a certain area is directly correlated with neural network-based predicting precision. The figure 9 and 10 presents the sample density and geo-related logarithmic prediction precision respectively in the test area, which also corresponds to the principle above, with high sample density and high precision in the center as well as low sample density and low precision at the edge.

To extend the generality of the neural network approach in this paper, we collect another scene dataset B, presented as figure 6(b), in which a blockage was added between the antenna and test area. In addition, datasets for beam prediction tasks from [39] are used to validate the generality of the proposed method as well, where the tasks focus on outdoor environments dominated by LOS components, rather than indoor environments with complex multipath conditions in this paper.

On the one hand, the proposed methods are trained on small scale datasets (i.e.,with 500 training samples) of several scenarios, shown as case 'B', 'scenario 6' and 'scenario 7'. The results indicate that the proposed methods are applicable to different environments and wireless settings, which promises good generality. Furthermore, the proposed method has better performance on scenarios from [39], attributed to plain LOS propagation conditions.

On the other hand, between the basically unobstructed dataset A and dataset B with blockage, as well as the multiple scenarios from [39], we explored ways to enhance the generality of the methods. When the training results from one scene are directly applied to other scenes for prediction, there is a significant performance degradation, as shown in figure 11 'B with A', '6 with 8', etc., which also implies to what degree these scenes are different from each other. However, it is possible to use transfer learning methods to exploit the image-channel knowledge of existing scenes [20], which can produce gains in training for new scenes. When there is already a large dataset of past scenes and a small dataset of new scenes, transfer learning is operated by firstly pre-training with the old dataset, and secondly fine tuned by the new dataset.

Such transfer learning has two advantages over re-training: it has a faster convergence speed, as shown in the cases 'A transfer to B' and '8 transfer to 7' of figure 12, and a higher prediction accuracy, as shown in the same cases of figure 11. However, certain scenarios may have few common features, such as '8 transfer to 6' case, so transfer learning does not provide significant improvement. To address this issue, data from more scenarios are used to build a richer historical experience. Here the datasets from scenario 5, 8 and 9 are combined as a more experienced dataset. It could been seen that case '5+8+9 transfer to 6' outperforms the individual training case 'scenario 6' and case '8 transfer to 6' in terms of convergence speed and prediction accuracy. In conclusion, transfer learning is regarded as a fast environment adaptation method, making this paper's approach applicable to more general and fast-changing environments. In the following, we use the datasets from this paper for further analysis, as these scenarios have more optimization space.

The limited feedback method with Grassmannian codebook [36], [38] is chosen as the wireless detection baseline. The Grassmannian codebook aims to find the most efficient way of quantizing complex vectors, so as to obtain a high SE

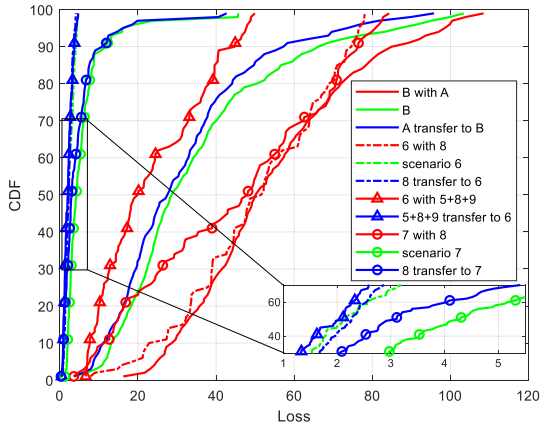


FIGURE 11. **Loss CDF of different training cases. Scenario 5-9 are from [39]. Legends for individual scenarios, such as 'scenario 6', means training on individual dataset of itself. 'A transfer to B' stands for transfer training from dataset A to dataset B. 'B with A' means implementing performance test on dataset B with the network trained on dataset A.**

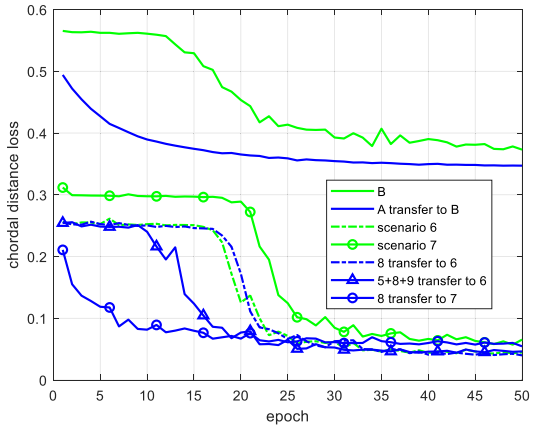


FIGURE 12. **Iterative loss for transfer learning.**

with limited feedback. It can be seen in figure 13 that the pure CV approach can outperform the wireless feedback approach with a certain overhead when it has an adequate dataset size, which itself does not require wireless overhead.

We then demonstrate the advantage of our straightforward approach compared to the beam classifier as in [11], [13], and [19]. Figure 13 shows that with growing dataset size, the average spectral efficiency increases, and as long as the dataset size is large enough, the CV method could approximate perfect CSI with any arbitrary precision. For 4 antenna configuration, the output of CNN is designed as a 8×1 vector, representing the real and imaginary part of the beam weight. Also a classifier for 8 beams could also be implemented as in [11], [13], and [19]. However, the targets of these methods are seeking the optimal codeword in a quantized codebook, which can not reach the perfect CSI upper bound.

Then the performance of fusion method is analyzed. The following figure 14 shows the average SE for several detection methods with different SNR. The 3-bit child codebook is

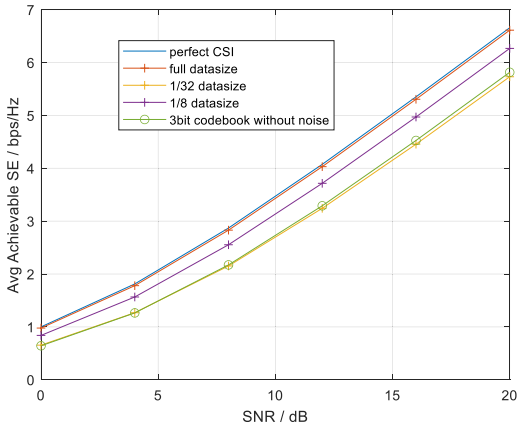


FIGURE 13. **Avg SE at different SNR with different dataset size.**

generated by a 4-bit parent codebook. An important opportunity of CV methods is that CV prediction performance is independent of wireless communication conditions Moreover, it is noticed that the combined methods are affected not only by SNR, but also by dataset size, which also leads out some new features.

It should be noted that thanks to the codeword rotation mechanism, the overhead of combined hierarchical feedback can be reduced, so that the fusion method can compete with the pure wireless mode under the same wireless overhead. Otherwise, notifying visual detection results will occupy most of the overhead, making it less competitive. For instance, the results of fusion method in figure 14 are based on maximum 3 bits wireless feedback. Without codeword rotation mechanism, additional 4 bits overhead of notifying the CV predicted parent codeword has to be used, increasing the overall overhead to 7 bits.

In order to research the effect of visual detection precision on the fusion methods, figure 15 is presented with different dataset sizes, which could be regarded as "SNR" for visual detection. In the context of this paper, the accuracy of beam prediction is discussed in two aspects: the accuracy of wireless detection and the accuracy of CNN prediction, especially for fusion methods, where both aspects contribute. On the one hand, the higher the SNR of the wireless link, the higher the wireless detection accuracy. On the other hand, when the neural network structure is fixed, the CNN prediction accuracy mainly depends on the dataset size; the larger the size, the higher the prediction accuracy. Therefore, the size of the dataset can be regarded as the "SNR" of the visual detection dimension, which determines the detection accuracy.

Further, while the performance of the fusion method can benefit from the accuracy improvement of both detection methods, it is in turn limited by the defects of both. It can be seen that at lower dataset size, the combined methods can obtain higher average achievable SE than pure wireless detection or CV detection. However, with dataset growing, the pure CV method gradually takes advantage and promises the best performance. So with limited wireless communication

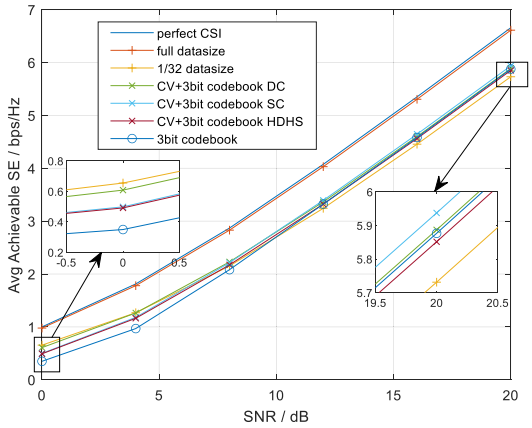
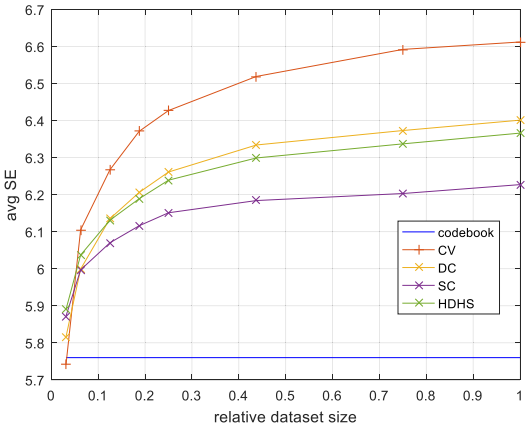
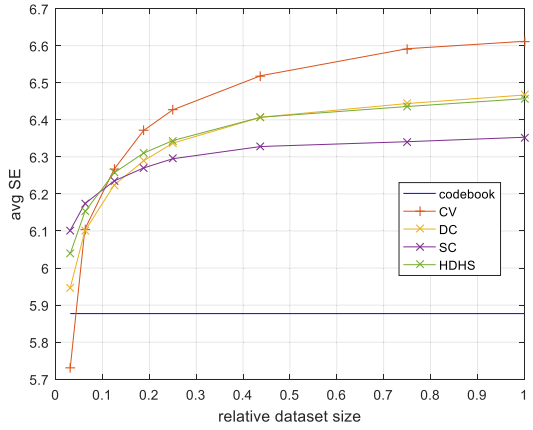


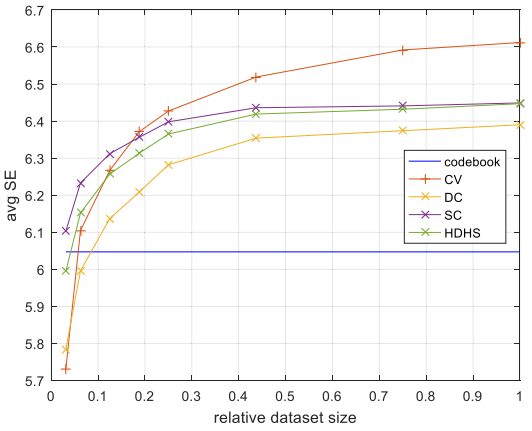
FIGURE 14. Avg SE at different SNR for different detection methods.



(a) Combined methods: CV+2bit child codebook generated by 4bit parent codebook, pure wireless: 2bit codebook, SNR=20dB



(b) Combined methods: CV+3bit child codebook generated by 4bit parent codebook, pure wireless: 3bit codebook, SNR=20dB



(c) Combined methods: CV+5bit child codebook generated by 6bit parent codebook, pure wireless: 3+2bit hierarchical codebook, SNR=20dB

FIGURE 15. Avg SE vs relative dataset size for different detection methods.

conditions, although the dataset could be large enough, the performance of combined methods are restricted by SNR. And more wireless feedback bits could also increase avg SE performed by combined methods, as well as enlarge the capable range for the combined methods to exceed pure visual detection.

To further clarify the capable range of combined methods, in other words, when it performs better than pure visual or wireless method, table 3 is presented. It can be seen that the fusion method outperforms individual CV or wireless detection under certain SNR and dataset size, even though it has the same feedback and pilot overhead with pure wireless detection.

It is noted that when either visual detection or wireless detection is ideal enough, wherein the former can be reflected in adequate dataset density and the latter in high-resolution codebook, this fusion method will lose competitiveness. It reveals that there is a matching issue between first-step visual detection accuracy and second-step wireless detection. On the one hand, if the visual detection accuracy is high, relatively low resolution hierarchical codebook will lower the detection accuracy in the second stage, making the prediction value that has converged to the best beam weight diverge again. This can be verified in table 1 that when dataset grows larger, pure CV method gradually takes advantage again. On the other hand, if the wireless detection accuracy is significantly higher than the visual detection accuracy, the resolution of the first stage will be difficult to match with the high-precision wireless feedback. At this time, the second-step codebook will not take effect in terms of narrowing the scope. This can be validated in table 1(b) and 1(c), wherein at 1/32 dataset size, CV+5bit configuration has little advantage compared to CV+3bit configuration and even perform worse at 8dB SNR.

Therefore, in order to ensure effectiveness or gain of the fusion method, it is necessary to match the resolution of visual detection and wireless sounding. Under the condition of limited wireless communication conditions, it is necessary to ensure that the visual detection accuracy can reach a certain threshold to achieve the effective gain of the fusion method, such as by improving the sampling density. In the case of limited sampling of visual dataset, it is also necessary to achieve a certain wireless feedback accuracy, such as a hierarchical

TABLE 3. The best detection methods at different visual and wireless detection conditions.

(A) COMBINED METHODS: CV+2BIT CHILD CODEBOOK GENERATED BY 4BIT PARENT CODEBOOK, PURE WIRELESS: 2BIT CODEBOOK

Avg. SE (bps/Hz)	1/32 dataset size	1/16 dataset size	1/8 dataset size	1/16 dataset size
SNR = 0 dB	CV 0.652	CV 0.783	CV 0.839	CV 0.879
SNR = 4 dB	CV 1.264	CV 1.477	CV 1.566	CV 1.629
SNR = 8 dB	CV 2.354	CV 2.435	CV 2.554	CV 2.637
SNR = 12 dB	**HDHS 3.362**	CV 3.576	CV 3.716	CV 3.811
SNR = 16 dB	**HDHS 4.601**	CV 4.816	CV 4.970	CV 5.072
SNR = 20 dB	**HDHS 5.890**	CV 6.105	CV 6.267	CV 6.371

(B) COMBINED METHODS: CV+3BIT CHILD CODEBOOK GENERATED BY 4BIT PARENT CODEBOOK, PURE WIRELESS: 3BIT CODEBOOK

Avg. SE (bps/Hz)	1/32 dataset size	1/16 dataset size	1/8 dataset size	1/16 dataset size
SNR = 0 dB	CV 0.652	CV 0.783	CV 0.839	CV 0.879
SNR = 4 dB	**SC 1.269**	CV 1.477	CV 1.566	CV 1.629
SNR = 8 dB	**SC 2.356**	CV 2.435	CV 2.554	CV 2.637
SNR = 12 dB	**SC 3.546**	**SC 3.611**	CV 3.716	CV 3.811
SNR = 16 dB	**SC 4.804**	**SC 4.875**	CV 4.970	CV 5.072
SNR = 20 dB	**SC 6.102**	**SC 6.174**	CV 6.267	CV 6.371

(C) COMBINED METHODS: CV+5BIT CHILD CODEBOOK GENERATED BY 6BIT PARENT CODEBOOK, PURE WIRELESS: 3+2BIT HIERARCHICAL CODEBOOK

Avg. SE (bps/Hz)	1/32 dataset size	1/16 dataset size	1/8 dataset size	1/16 dataset size
SNR = 0 dB	CV 0.652	CV 0.783	CV 0.839	CV 0.879
SNR = 4 dB	CV 1.264	CV 1.477	CV 1.566	CV 1.629
SNR = 8 dB	CV 2.354	**SC 2.461**	CV 2.554	CV 2.637
SNR = 12 dB	**SC 3.55**	**SC 3.665**	**SC 3.74**	CV 3.811
SNR = 16 dB	**SC 4.806**	**SC 4.932**	**SC 5.01**	CV 5.072
SNR = 20 dB	**SC 6.103**	**SC 6.232**	**SC 6.312**	CV 6.371

codebook with an appropriate codeword distance. Because this paper uses typical wireless hierarchical codebook as the second-step codebook of the fusion feedback method, but does not customize the codebook according to the visual accuracy, so there are still limitations on performance. For example, when both the codebook resolution and sampling density are high, the fusion method cannot exceed the two exclusive methods. This is because it is difficult to generate high-resolution Grassmannian codebook, and it is also impossible to obtain an ideal child codebook for the method in this paper. This problem may also need to be solved with customized codebook, which remains to be studied. However, although there is no special customization, the fusion method in this paper is verified to be able to have better performance than exclusive visual or wireless detection method.

Regarding the energy consumption considerations of the methods in this paper, on the one hand, since there are no strict requirements for deployment locations, camera sensors deployed in infrastructures [2] can also be utilized for the visual detection purposes of this paper, without additional deployment costs or energy consumption. On the other hand, even if exclusive camera devices are used, the method of this paper requires high frame rate but low resolution image flow, which could be implemented with low energy consumption [40]. In addition, high frame rate cameras could be woken up or put to sleep by other lower-power cameras that monitor whether targets enter the field of view [41]. In this way, the required average energy consumption could be greatly reduced.

VI. CONCLUSION

In this paper, we propose a CV-aided straightforward beam weight prediction method. In order to improve the weight prediction resolution of visual detection, a fusion method of CV detection and limited wireless feedback has been proposed. Moreover, this method utilizes a codeword rotation mechanism implemented by Householder transform to save the notification overhead of CV prediction results. A testbed has been built to gather joint visual-wireless dataset. Experimental results show that the proposed straightforward prediction method is able to map actual scene image to optimal beam weights. And the fusion method is verified to perform better than exclusive visual or wireless detection, under the condition of using appropriate child codebook at certain SNR and CV sampling density.

REFERENCES

[1] J. G. Andrews et al., "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014, doi: 10.1109/JSAC.2014.2328098.

[2] A. Ali, N. Gonzalez-Prelcic, R. W. Heath Jr., and A. Ghosh, "Leveraging sensing at the infrastructure for mmWave communication," *IEEE Commun. Mag.*, vol. 58, no. 7, pp. 84–89, Jul. 2020, doi: 10.1109/MCOM.001.1900700.

[3] Y. Xie and P. Ren, "Reliability-optimal pilot-assisted transmission for URLLC over non-reciprocal MISO channels: TDD or FDD?" *China Commun.*, vol. 19, no. 4, pp. 14–27, Apr. 2022, doi: 10.23919/JCC.2022.04.002.

[4] S. Han, I. Chih-Lin, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015, doi: 10.1109/MCOM.2015.7010533.

[5] T. Xiang, Y. Wang, H. Li, B. Guo, and X. Zhang, "A computer vision based beamforming scheme for millimeter wave communication in LOS scenarios," in *Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT)*, Oct. 2019, pp. 401–407, doi: 10.1109/ICCSNT47585.2019.8962465.

[6] T. Xiang, H. Li, B. Guo, and X. Zhang, "A computer vision aided beamforming scheme with EM exposure control in outdoor LOS scenarios," in *Proc. IEEE 3rd Int. Conf. Inf. Commun. Signal Process. (ICICSP)*, Sep. 2020, pp. 241–246, doi: 10.1109/ICICSP50920.2020.9232040.

[7] T. Xiang, J. Gu, B. Yu, and X. Zhang, "A sensing-based over-the-air phased array antennas phase calibration method with greedy algorithm," *IEEE Antennas Wireless Propag. Lett.*, vol. 20, no. 12, pp. 2240–2244, Dec. 2021, doi: 10.1109/LAWP.2021.3104611.

[8] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, "ViWi: A deep learning dataset framework for vision-aided wireless communications," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5, doi: 10.1109/VTC2020-Spring48590.2020.9128579.

[9] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided dynamic blockage prediction for 6G wireless communication networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–6, doi: 10.1109/ICCWorkshops50388.2021.9473651.

[10] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided 6G wireless communications: Blockage prediction and proactive handoff," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10193–10208, Oct. 2021, doi: 10.1109/TVT.2021.3104219.

[11] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5, doi: 10.1109/VTC2020-Spring48590.2020.9129369.

[12] H. Ahn et al., "Machine learning-based vision-aided beam selection for mmWave multiuser MISO system," *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1263–1267, Jun. 2022, doi: 10.1109/LWC.2022.3163780.

[13] T. Zhang, J. Liu, and F. Gao, "Vision aided beam tracking and frequency handoff for mmWave communications," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2022, pp. 1–2, doi: 10.1109/INFOCOMWKSHPS54753.2022.9798197.

[14] M. Zecchin, M. B. Mashhadi, M. Jankowski, D. Gündüz, M. Kountouris, and D. Gesbert, "LiDAR and position-aided mmWave beam selection with non-local CNNs and curriculum training," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 2979–2990, Mar. 2022, doi: 10.1109/TVT.2022.3142513.

[15] F. Talaei and X. Dong, "Hybrid mmWave MIMO-OFDM channel estimation based on the multi-band sparse structure of channel," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1018–1030, Feb. 2019, doi: 10.1109/TCOMM.2018.2871448.

[16] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[17] M. Alrabeiah and A. Alkhateeb, "Deep learning for TDD and FDD massive MIMO: Mapping channels in space and frequency," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019, pp. 1465–1470, doi: 10.1109/IEEECONF44664.2019.9048929.

[18] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, Sep. 2020, doi: 10.1109/TCOMM.2020.3003670.

[19] F. Gao, B. Lin, C. Bian, T. Zhou, J. Qian, and H. Wang, "FusionNet: Enhanced beam prediction for mmWave communications using sub-6 GHz channel and a few pilots," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8488–8500, Dec. 2021, doi: 10.1109/TCOMM.2021.3110301.

[20] Y. Yang, F. Gao, Z. Zhong, B. Ai, and A. Alkhateeb, "Deep transfer learning-based downlink channel prediction for FDD massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7485–7497, Dec. 2020, doi: 10.1109/TCOMM.2020.3019077.

[21] W. Xu, F. Gao, J. Zhang, X. Tao, A. Alkhateeb, and S. Ma, "Deep learning based channel covariance matrix estimation with scene images," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Jul. 2021, pp. 162–166, doi: 10.1109/ICCC52777.2021.9580233.

[22] D. Liu, W. Ma, S. Shao, Y. Shen, and Y. Tang, "Performance analysis of TDD reciprocity calibration for massive MU-MIMO systems with ZF beamforming," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 113–116, Jan. 2016, doi: 10.1109/LCOMM.2015.2499283.

[23] J. Cardillo and M. A. Sid-Ahmed, "3-D position sensing using a passive monocular vision system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 809–813, Aug. 1991, doi: 10.1109/34.85671.

[24] B. Yu and X. Zhang, "A small range ergodic beamforming method based on binocular vision positioning," in *Proc. 7th IEEE Int. Conf. Netw. Intell. Digit. Content (IC-NIDC)*, Nov. 2021, pp. 168–171, doi: 10.1109/IC-NIDC54101.2021.9660448.

[25] X. Chen, Z. Wei, X. Zhang, and L. Sang, "A beamforming method based on image tracking and positioning in the LOS scenario," in *Proc. IEEE 17th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2017, pp. 1628–1633, doi: 10.1109/ICCT.2017.8359906.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[27] *CV Aided Beamforming Fused With Limited Feedback Dataset*. Accessed: Dec. 14, 2022. [Online]. Available: https://github.com/Mrxtq/CV-aided-beamforming-fused-with-limited-feedback-dataset

[28] D. J. Love, R. W. Heath Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003, doi: 10.1109/TIT.2003.817466.

[29] B. Tahir, S. Schwarz, and M. Rupp, "Constructing Grassmannian frames by an iterative collision-based packing," *IEEE Signal Process. Lett.*, vol. 26, no. 7, pp. 1056–1060, Jul. 2019, doi: 10.1109/LSP.2019.2919391.

[30] K. Chen and C. Qi, "Beam training based on dynamic hierarchical codebook for millimeter wave massive MIMO," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 132–135, Jan. 2019, doi: 10.1109/LCOMM.2018.2881084.

[31] K. Ko and J. Lee, "Regenerative hierarchical codebooks for limited channel feedback in MIMO systems," in *Proc. Inf. Theory Appl. Workshop*, Feb. 2012, pp. 398–400, doi: 10.1109/ITA.2012.6181770.

[32] M. Egan, C. K. Sung, and I. B. Collings, "Structured and sparse limited feedback codebooks for multiuser MIMO," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 3710–3721, Aug. 2013, doi: 10.1109/TWC.2013.051013.120782.

[33] J. Yuan, X. Li, W. Pu, and R. Yuan, "A two-stage codebook design based on beam perturbation for massive MIMO systems," in *Proc. 31st Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA)*, Mar. 2017, pp. 29–34, doi: 10.1109/WAINA.2017.35.

[34] H. E. A. Laue and W. P. D. Plessis, "A coherence-based algorithm for optimizing rank-1 Grassmannian codebooks," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 823–827, Jun. 2017, doi: 10.1109/LSP.2017.2690466.

[35] S. Schwarz, M. Rupp, and S. Wesemann, "Grassmannian product codebooks for limited feedback massive MIMO with two-tier precoding," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 1119–1135, Sep. 2019, doi: 10.1109/JSTSP.2019.2930890.

[36] (Jan. 20, 1997). *Packings in Grassmannian Spaces*. Accessed: Dec. 23, 2022. [Online]. Available: https://neilsloane.com

[37] K. Ko and J. Lee, "Hierarchical codebook design for fast search with Grassmannian codebook," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2012, pp. 873–877, doi: 10.1109/WCNC.2012.6214496.

[38] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016, doi: 10.1109/TWC.2016.2520930.

[39] A. Alkhateeb et al., "DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 122–128, Sep. 2023, doi: 10.1109/MCOM.006.2200730.

[40] A. Rowe et al., "CMUcam3: An open programmable embedded vision sensor," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. RITR-07-13, 2007.

[41] S. Hengstler, D. Prashanth, S. Fong, and H. Aghajan, "MeshEye: A hybrid-resolution smart camera mote for applications in distributed intelligent surveillance," in *Proc. 6th Int. Symp. Inf. Process. Sensor Netw.*, Cambridge, MA, USA, Apr. 2007, pp. 360–369, doi: 10.1109/IPSN.2007.4379696.