Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4490790, 14 pages
https://doi.org/10.1155/2022/4490790

Review Article

A Survey of Compressive Data Gathering in WSNs for IoTs

Xun Wang and Hongbin Chen

Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing, Guilin University of Electronic Technology, Guilin 541004, China

Correspondence should be addressed to Hongbin Chen; chbscut@guet.edu.cn

Received 20 October 2021; Accepted 28 December 2021; Published 25 January 2022

Academic Editor: Manuel Fernandez-Veiga

Internet of Things (IoTs) are increasingly widespread in the field of health care, smart city and smart home application, industrial and agricultural monitoring, automation, etc. With its growing scale of networks, there are a large amount of data in IoTs needing to be sensed, transmitted, and processed. Resource-limited Wireless Sensor Networks (WSNs) as a perceptual layer of IoTs are hard to handle massive uncompressed sensing data. Compressive data gathering (CDG), which applies compressive sensing theory to data gathering, is a perfectly matching method for data compressing and gathering in WSNs. This promising method has attracted lots of researchers' attention. In this paper, we attempt to survey substantial references about CDG in WSNs. According to their technology schemes, we classify the published references into three categories, i.e., routing protocol of CDG, clustering scheme of CDG, and CDG combined with other technologies. The merits and defects of each method are highlighted. Our work aims to provide an insight into CDG and promote improvements of this technology.

1. Introduction

The expansion of Internet of Things (IoTs) has given rise to smart city, smart home, industrial and agricultural monitoring and automation, intelligent transportation, and so on, which immensely facilitates human's life. As the perceptual layer of IoTs, Wireless Sensor Networks (WSNs) are responsible for data sensing and transmitting to upper layer for further decision [1]. A WSN is composed of hundreds to thousands of sensor nodes acquiring environmental parameters such as temperature, humidity, flow, distance, and velocity, and then, these data are transmitted to the sink or fusion center for further processing. So data gathering is a crucial function in WSNs.

Generally speaking, sensor nodes are battery-powered and cannot be recharged normally. Once nodes' energy is depleted, a WSN is out of work. So energy efficiency is the most important consideration in WSNs. Energy in a WSN is mainly consumed in three aspects: sensing, data processing, and communication. Moreover, data communication expends much more energy than the other two aspects. Nodes periodically sense environmental data and transmit them to the sink. The amount of data transmission is huge and consumes a large amount of energy. So we need a more energy-efficient method for data gathering in WSNs.

Data compression technology is usually applied in data gathering in WSNs. However, traditional data compression methods such as entropy coding often bring significant extra computation and control overheads which are not suitable for nodes with limited computing resources. Compressive sensing (CS) theory was introduced in data gathering, denoted as compressive data gathering (CDG) or compressive data aggregation (we use uniform abbreviation of CDG in this paper), to reduce global data traffic. Besides, in the encoding process of CDG, only multiplication and addition need to be operated by nodes, which is more suitable for WSNs with limited computational resources. The complicated data recovery algorithm is performed at the sink which is regarded as having unlimited resources. Furthermore, compressing and sampling are implemented simultaneously in CDG, which can save storage resources in WSNs.

Owing to the advantages of CDG, in recent years, researchers have extended a large amount of study on

CDG. Reference [2] proposed a complete scheme to apply CS to data gathering in large-scale WSNs for the first time, which confirmed the advantages of CDG. The main objectives of CDG studies are centered on, with the constraint of data recovery accuracy, reducing the number of data transmission and then cutting down the network energy consumption. At the meantime, the traffic load balance needs to be considered. Eventually, the lifetime of WSNs can be prolonged.

Around these goals, there are a lot of CDG schemes proposed in references. We classify these researches as follows. Firstly, we focus on the routing protocol of CDG because routing protocol is essential in data gathering methods. So there are a lot of researches studying more efficient routing schemes for CDG in references. Secondly, we summarize clustering schemes of CDG. Clustering topology control technology is widely used in WSNs for load balance and energy efficiency. So it can be applied in CDG for performance enhancement. At last, we notice that some other technologies, such as mobile collector, sleep scheduling, and heuristic and learning algorithms, are applied in CDG for optimizing energy efficiency and longevity of WSNs.

In this paper, we commit to giving a comprehensive introduction about CDG based on published literature in the past decade. We hope that through this work, the reader can gain an insight into the promising CDG technology and explore a wide application prospect. The rest of this paper is organized as follows. Section 2 presents some preliminaries about CDG, such as general network models, fundamental of CS theory, and main problems in CDG. Section 3 and Section 4 give detailed descriptions of CDG methods in the above categories, respectively. Finally, some concluding remarks are made in Section 6.

2. System Model and Problem Formulation

2.1. Network Model. Generally, a WSN of N sensor nodes is deployed randomly in a regular or irregular area. The distribution of nodes can be either even or uneven. All the nodes send their sensing data to the sink through directed link or multihops. The sink is located at the center or edge of the area. In some particular situations, the location of nodes needs to be acquired. The network is self-organized, and data rates of all nodes can be equal or unequal. The network routing protocol can be chain-type, tree-type, or others. Topology control technology such as clustering can usually be adopted in the network.

2.2. Compressed Sensing Preliminaries. According to the CS theory [3], a sparse or compressible signal can be reconstructed from far fewer samples (measurements) than that of the traditional Nyquist formula required. If a signal $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T, \mathbf{x} \in \mathbf{R}^{N\times 1}$, or its transform $\mathbf{s}$ under some orthogonal basis $\mathbf{\Psi} = [\psi_1, \psi_2, \cdots, \psi_N]^T, \mathbf{\Psi} \in \mathbf{R}^{N\times N}$, i.e., $\mathbf{x} = \mathbf{\Psi s}$, has only k nonzero elements, it can be defined as k-sparse signal. $\mathbf{\Psi}$ is called the sparse basis.

Then, we can compress this k-sparse signal $\mathbf{x}$ into a vector $\mathbf{y}$ with M measurements using CS theory, which is also referred to as $\mathbf{y}$ is the projection of $\mathbf{x}$. That is, $\mathbf{y} = \mathbf{\Phi x} = \mathbf{\Phi\Psi s}$, where $\mathbf{\Phi} = \{\boldsymbol{\varphi}_{ij}\}$ is a $M \times N$ matrix called the measurement matrix or projection matrix. We also denote $\mathbf{\Theta} = \mathbf{\Phi\Psi}$, and $\mathbf{\Theta}$ is the sensing matrix. If the measurement matrix $\mathbf{\Phi}$ meets the Restricted Isometry Property (RIP) condition [4], or $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are uncorrelated, we can recover the original signal $\mathbf{x}$ from M measurements in $\mathbf{y}$. $M \geq C \cdot k \cdot \log N$, where C is a small constant. Generally, $M << N$ and the dimension of $\mathbf{x}$ is compressed from N to M.

The independently and identically distributed (i.i.d.) Gaussian random measurement matrix and Bernoulli random matrix with a uniform number of ± 1 are universal CS measurement matrices [3]. Meanwhile, Fourier, discrete cosine, and wavelet transform basis are commonly used as the sparse basis $\mathbf{\Psi}$.

To recover $\mathbf{x}$ from $\mathbf{y}$, the following convex optimization problem needs to be solved:

$$\min_{\mathbf{s}\in\mathbf{R}^N} \|\mathbf{s}\|_{l_1} \text{ subject to } \mathbf{y} = \mathbf{\Phi\Psi s}. \tag{1}$$

There are commonly used reconstruction algorithms such as Basis Pursuit (BP) algorithm and Orthogonal Matching Pursuit (OMP) greedy algorithm [5].

The signal $\mathbf{x}$ can be regarded as the data vector sampled by N sensor nodes in the WSN. The primary mission of data gathering is that collecting the data vector $\mathbf{x}$ and sending them to the sink as effectively as possible. Hence, with CDG, we can collect vector $\mathbf{y}$ with M measurements instead of original $\mathbf{x}$, by which data transmissions in the network are greatly cut down.

2.3. Problem Formulation. The main objectives of CDG are reducing data transmission, load balance, thereby improving energy efficiency and prolonging lifetime of WSNs for IoTs. Besides, accuracy of recovery data, the networks' scalability, and robustness of the CDG algorithms are also considered in some references.

To achieve the above goals, proper routing and topology control scheme matching CDG process should be specially designed with consideration of energy consumption and load balance, etc. At the meantime, the process of data compression is dependent with the CS measurement matrix $\mathbf{\Phi}$. So constructing the optimal measurement matrix to reduce data transmissions and promote accuracy of recovery data is also a valuable research direction. Last but not the least, with the advance of CDG research, scholars incorporate multiple methods in CDG and hope to build green and intelligent WSNs for IoTs.

The CDG methods are evaluated mainly in terms of energy consumption, lifetime, reconstruction accuracy, and system robustness to node failures.

3. Routing Protocol of CDG

The realization of CDG and design of energy-efficient routing protocols in WSNs are the primary task in early research work. In this section, we will introduce three types of basic implementation frameworks firstly. And then, two categories of route topologies, i.e., determined routing and random

Table 1: Routing protocols of CDG.

Ref.	CDG scheme	Routing method	Main contribution	Year
[8]	Sparse CDG	MST	Propose a distributed method of sparse CDG	2014
[9]	Sparse CDG	eMSTP	Take the sink as the root of MST in sparse CDG	2014
[10]	Sparse CDG	WCDA	Consider the tree's cost which taking hops and distances in routing paths	2016
[11]	Sparse CDG	Random walk (RW)	Firstly adopt random walk (RW) in sparse CDG	2015
[12]	Sparse CDG	RW	Analyze the optimal transmission range of nodes in RW-based CDG	2017
[13]	Sparse CDG	RW	Propose a global mathematical model of RW-based CDG	2019
[14]	Sparse CDG	Direct RW	Overcome the weakness of directionless feature in RW-based CDG	2019

walk (RW) routing, are detailed. We present an overview of routing protocols of CDG in Table 1.

3.1. Three Types of Basic CDG Frameworks. Luo et al. proposed a complete scheme to realize CDG in large-scale WSNs for the first time in [2], which becomes one of basic frameworks for CDG. In CDG, the ultimate goal is receiving a vector $\mathbf{y}$ with M measurements at the sink. So in [2], in a chain-type routing scheme, as in Figure 1(a), each node s_j in the network firstly generates M random coefficients $\boldsymbol{\varphi}_{ij} = \{\varphi_{1j}, \varphi_{2j}, \cdots, \varphi_{Mj}\}^T$, $i = 1, 2 \cdots, M$, $j = 1, 2 \cdots, N$, where T represents the transposition of the vector. The sensing data of s_j is denoted as d_j. Then, the initial node s_1 multiplies its sensing data d_1 to $\boldsymbol{\varphi}_{i1}$ and sends the product $\boldsymbol{\varphi}_{i1}d_1$ to the next node s_2. Node s_2 adds its own product to $\boldsymbol{\varphi}_{i1}d_1$ and gets $\boldsymbol{\varphi}_{i1}d_1 + \boldsymbol{\varphi}_{i2}d_2$ which will be transmitted to node s_3. The same procedure is repeated one by one along the chain routing path until to the sink. At last, the sink will receive the following vector $\mathbf{y}$:

$$\begin{aligned}\mathbf{y} &= \boldsymbol{\Phi}\mathbf{d} = \boldsymbol{\varphi}_{i1}d_1 + \boldsymbol{\varphi}_{i2}d_2 + \cdots + \boldsymbol{\varphi}_{\mathrm{iN}}d_N \\ &= [\boldsymbol{\varphi}_{i1}, \boldsymbol{\varphi}_{i2}, \cdots, \boldsymbol{\varphi}_{\mathrm{iN}}][d_1, d_2, \cdots, d_N]^T \\ &= \begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1N} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2N} \\ \vdots & \vdots & \vdots & \\ \varphi_{M1} & \varphi_{M2} & \cdots & \varphi_{MN} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{N} \varphi_{1j}d_j \\ \sum_{j=1}^{N} \varphi_{2j}d_j \\ \vdots \\ \sum_{j=1}^{N} \varphi_{Mj}d_j \end{bmatrix},\end{aligned} \tag{2}$$

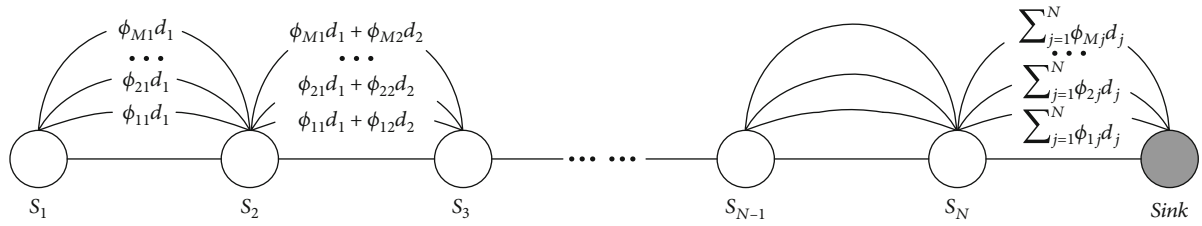
where $\mathbf{d} = [d_1 d_2 \cdots d_N]^T$.

The procedure of CDG in a tree-type routing scheme is similar to that in a chain-type routing scheme. In this scheme, all nodes transfer the same size of the data packet no matter their locations in routing paths, which improves load balance. So the total number of transmissions of the above network is MN. Compared with the baseline data gathering without CS in Figure 1(b), the total number of transmissions is up to $O(N^2)$. When $M << N$, the data traffic and the energy consumption of the network can be significantly cut down.

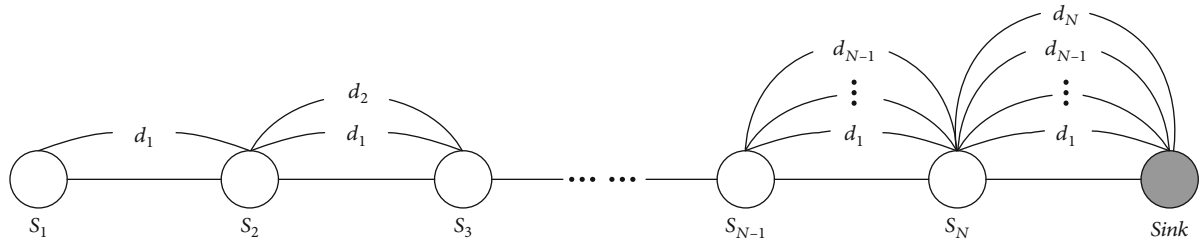
However, in [6], Luo et al. pointed out that in Figure 1(a), nodes $s_1, s_2 \cdots s_{M-1}$ transmit redundant data because their data traffic is even higher than in baseline data gathering in Figure 1(b). If the sparsity k of the sensing data is large, the required number M of measurements is large and redundant data will be increased which results in a waste of nodes' energy. So they improve the measurement matrix $\boldsymbol{\Phi}$ as $\boldsymbol{\Phi} = [\mathbf{I}\,\mathbf{R}]$, where $\mathbf{I}$ is a $M \times M$ identity matrix and $\mathbf{R}$ is a $M \times (N - M)$ fully random Gaussian matrix. Using the improved measurement matrix $[\mathbf{I}\,\mathbf{R}]$, the first M nodes simply send their original sensor data to node s_{M+1}. It is also proved in [6] that the matrix $[\mathbf{I}\,\mathbf{R}]$ satisfies RIP for successful CS recovery. Reference [7] denoted CDG in [2] as plain CDG and designed a hybrid CDG to address the aforementioned problem of redundant data in plain CDG. In hybrid CDG, the node only applies CDG in the case when the number of transmitted data packets is more than the required number of measurements M. If the number of forwarded data is less than M, the node forwards the original data without using CS. In Figure 2, we summarize the three data-gathering schemes, i.e., non-CDG, plain CDG, and hybrid CDG, in the same subtree of a network for better comparison. The digits next to transmission links represent the size of data packets along this link. In hybrid CDG as shown in Figure 2(c), assuming $M = 3$, nodes 7, 11, and 12 adopt CDG and the other nodes directly send their sensing data to their parents. While in plain CDG as shown in Figure 2(b), each node needs to send three data packets. Obviously, data traffic in hybrid CDG is smaller than in plain CDG. Thus, hybrid CDG is the commonly used CDG method in other references.

Nevertheless, both plain CDG and hybrid CDG require all the nodes to participate in measurement gathering, in which the projection matrix $\boldsymbol{\Phi}$ is dense. In [8], Wang et al. proved that a compressible data vector could be recovered from its sparse random projections. Sparse random projections can be acquired through a sparse random projection matrix $\boldsymbol{\Phi}$ which contains entries

$$\varphi_{ij} = \sqrt{s} \begin{cases} +1 & \text{with prob. } \dfrac{1}{2s}, \\ 0 & \text{with prob. } 1 - \dfrac{1}{s}, \\ -1 & \text{with prob. } \dfrac{1}{2s}, \end{cases} \tag{3}$$



(a) CDG in chain-type routing



(b) Baseline data gathering

Figure 1

where s is a constant to control the probability. Each φ_{ij} equals nonzero with the probability $1/s$. The work in [8] lays the theoretical foundation for sparse CDG, i.e., a CDG scheme that does not need all nodes to send their data to form CS measurements. The fewer nodes take part in CDG; the lower energy is consumed in the network.

Reference [10] constructed a sparse random projection matrix $\mathbf{\Phi}_{m\times n}$ for CDG. There are $\lceil n/m \rceil$ nonzero elements in each row of $\mathbf{\Phi}_{m\times n}$, where n and m represent the number of nodes and required number of measurements, respectively. Meanwhile, in order to recover the original data successfully, it is required that none of the columns in $\mathbf{\Phi}_{m\times n}$ contains all zero entries. They took $\mathbf{\Phi}_{m\times n}$ as the measurement matrix and designed a sparse CDG scheme. Firstly, m nodes are selected randomly as projection nodes. Each projection node i is allocated a coefficient vector $\boldsymbol{\varphi}_{ij} = \{\varphi_{i1}, \varphi_{i2}, \cdots, \varphi_{\text{in}}\}$ which is used to collect a CS measurement, where $i = 1, 2, \cdots, m, j = 1, 2, \cdots, n$, and $\boldsymbol{\varphi}_{ij}$ are a row from $\mathbf{\Phi}_{m\times n}$. Each coefficient φ_{ij} in $\boldsymbol{\varphi}_{ij}$ corresponds to a node j in the WSN. The nodes with coefficients $\varphi_{ij} \neq 0$ are called interest nodes of node i. The nodes with coefficients $\varphi_{ij} = 0$ do not send data. A Minimum-Spanning-Tree (MST) is built for each projection node i connecting all of its interest nodes. Then, the projection node i receives data from its interest nodes and computes the weighted sum $\mathbf{y_i} = \sum_{j=1}^{n}\varphi_{ij}d_j$, that is, one of the m measurements in $\mathbf{y}$. And then, node i sends $\mathbf{y_i}$ to the sink through the shortest path. In this sparse CDG scheme, partial nodes in WSNs need to take part in data gathering, which cut down data transmissions and energy consumption in the network.

Though the above researches are incipient work in the early stage, they paved the way for further study of CDG in WSNs for IoTs. Plain CDG, hybrid CDG, and sparse CDG are basic frameworks in subsequent research literature.

3.2. Determined Routing Schemes of CDG. In the above sparse CDG method, exchange information between projection nodes and interest nodes consumes a lot of communication resources. So the authors improved the above sparse CDG method in [10] and adopted a distributed method in sparse CDG. In this distributed algorithm, initially, the sink sends a discovery message to its neighbors, and then, these neighbors broadcast their received message to other nodes far away from the sink. In this way, each node can obtain its shortest path to the sink and hops along the shortest path. The measurement matrix $\mathbf{\Phi}$ is stored in nodes' memory. Thus, any node can check whether its neighbor nodes belong to the interest node set of a certain routing tree. Depending on this information, nodes can locally decide to construct and maintain the forward tree. Distributed methods do not require complete topological information of the network, avoiding a large amount of signaling overhead.

Reference [11] summarized research work in [9, 10] and proposed the Minimum Spanning Tree Projection (MSTP) algorithm as well as improved the MSTP algorithm (eMSTP). The eMSTP algorithm takes the sink into MST as the root of the tree, which can save transmission overhead from projection nodes to the sink. MSTP and eMSTP are compared with non-CDG, plain CDG, and hybrid CDG. Simulation results showed that the proposed algorithms are superior in terms of network throughput, load balance, and network lifetime.

However, the MST used in the above methods merely considers least hops from interest nodes to projection nodes when constituting the routing tree. Least hops do not mean shortest distance. According to the energy consumption model of wireless communication, the energy consumption is proportional to transmission distance. So Abbasi-Daresari and Abouei focused on the load balance issue and proposed a Weighted Compressive Data Aggregation (WCDA) method in [12]. The sparse random measurement matrix is also used in WCDA, i.e., WCDA is also a sparse CDG scheme. The main innovation in WCDA is that both hops and distance from candidate nodes (the same as interest nodes) to collector nodes (the same as projection nodes)

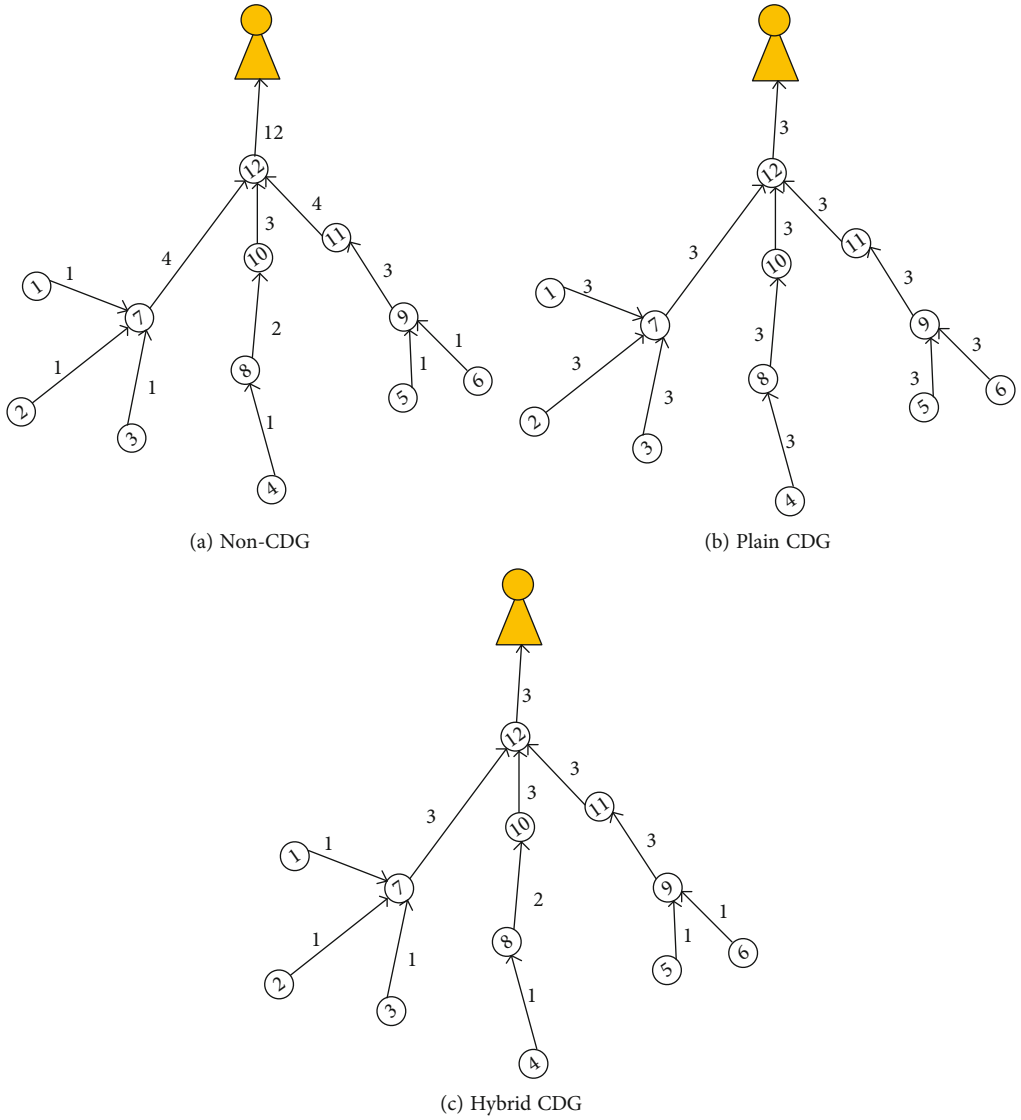


Figure 2: The three types of CDG.

are considered into a performance metric termed as tree's cost. Candidate nodes are connected to the routing tree through the link with the minimum tree's cost. So routing paths in WCDA are more energy-efficient. Simulation results indicate that WCDA can save 20% energy consumption and reduce 10% load variance compared with MSTP in [11].

However, merely improving the routing protocol of CDG is far from enough.

3.3. Random Walk Routing of CDG. The route protocols in Subsections 3.1 and 3.2 belong to the determined route scheme. However, a random walk (RW) routing scheme is also commonly used in WSNs for IoTs. Compared to the determined route protocol, RW does not require computing routing paths precisely, which can save computing resources and extra information exchanges and enhance network dynamic adaptability. Furthermore, RW can avoid attacks and reach better load balance due to its undetermined routing path. So CDG methods based on RW are studied in several references.

In [13], Zheng et al. made the first attempt to adopt RW in CDG. The m CS measurements are acquired through m random walks. First, m nodes are selected randomly to initialize m independent random walks of length t. Then, each of these initial nodes chose one of its neighbors to send its weighted sensing data. At the end of each walk, the last nodes will obtain one random projection and transmit it to the sink through the shortest path. They prove that $m = O(k \log(n/k))$ independent random walks with the length of each walk $t = O(n/k)$ are sufficient for recovering a k-sparse signal in WSNs with n nodes. The m random walks can also be performed simultaneously. Simulation results in [13] show that the proposed scheme can significantly reduce the communication cost compared to CDG schemes using a determined route.

After the breakthrough in RW-based CDG in [13], the subsequent researches attempt to do some optimal works.

Reference [14] integrated CDG with RW (denoted as CSR) for the purpose of saving energy. They study the trade-off between the nodes' transmission range R and the RW length to suggest an optimal transmission range $R*$ for the least energy consumption of the network. Two models of transmitting CS measurements to the sink, i.e., D-CSR and M-CSR in which CS measurements are transmitted directly and through multi-hop, respectively, are compared in order to find the optimal case that consumes the least energy. Both theoretical analysis and simulation results show that M-CSR can reduce up to 30% energy consumption compared to D-CSR.

Reference [15] modeled the process of RW-based CDG as an absorbing Markov chain. The sink node is regarded as the absorbing state. The transition probability corresponds to the probability that a sensor node takes part in RWs. The expectation of energy cost is used as the energy consumption model of the network. The optimal transition probability to minimize the expected energy cost of the network was analyzed. The work in [15] provides a global mathematical method to study RW-based CDG problems

However, in the above schemes, RW is directionless which may be routed to any node in the network, sometimes may be far away from the sink. It will increase communication cost. Static routing paths are still needed, which are used to transmit CS measurements to the sink. That would degrade dynamic adaptability of networks. Reference [16] proposed a dual RW-based CDG method to address these problems. The dual RW includes a directed RW and a same layer RW. The directed RW starts at any node and ends at the sink to collect CS measurements, which avoids a static route used to transport CS measurements to the sink and save the communication cost. Since all directed RW paths terminate at the sink, it will lead to a spatial nonuniform sample. So the same layer RW is used to compensate for the spatial nonuniform sample. The proposed scheme does not rely on coordinate information of nodes. A ring hierarchical network topology is built according to nodes' relative distance from the sink. The neighbors of each node are divided into neighbor sets of current layer and upper layer, respectively. The directed RW is based on neighbor sets of upper layer, and the same layer RW is based on neighbor sets of current layer. These two types of RW connect each other with a certain probability.

The above work in this subsection combines merits of both RW and CDG to improve network energy efficiency and dynamic adaptability and get preferable results. However, randomness is a double-edged sword. Despite that RW can save computational resources of the network, it may bring some extra controlling overhead. Moreover, the accuracy of recovery data is not mentioned compared with determined route CDG methods.

4. Clustering Schemes of CDG

The work in Section 3 only considers route design of CDG, which has limited effect on enhancing performances of WSNs for IoTs. It is verified that topology control technologies, such as clustering technique, are beneficial for traffic load balance and energy efficiency of the network, especially in large-scale WSNs. So there is a large amount of literature devoted to study the clustering method in CDG, which is denoted as clustering CDG.

Generally speaking, the problems in a clustering CDG include determining the number of clusters, CH selection, cluster formation, and scheme of data transmission. The objectives still center on reducing data transmission, enhancing energy efficiency, balancing traffic load, and at last prolonging the lifetime of the network. Different literature may aim at one or more problems to conduct a study. We present an overview of clustering schemes of CDG in Table 2.

Xie and Jia proposed a transmission-efficient clustering algorithm using hybrid CDG in [17]. It is the earlier study about a clustering method in CDG. The objective of the proposed algorithm is to minimize the total number of transmissions of the network. They analyze the optimal size N_c^* of clusters achieving minimum number of transmissions based on an analytical model. Thus, the optimal number of clusters in the network with N nodes is $C = N/N_c^*$. Meanwhile, the distribution of CH is determined by solving a k-median problem that is to find the location of C CHs in the network such that the total hops from all cluster members (CMs) to their nearest CHs is minimized. When clusters in the network are formed, data transmission is initiated. Data transmission is split into two levels: intracluster and intercluster. In the intracluster transmission phase, CMs send data to the cluster head (CH) through shortest path without CDG. While in the intercluster transmission phase, CHs compressed CMs' data using hybrid CDG and then send data to the sink through the backbone tree. It is a widely used transmission scheme of the clustering CDG in later literature.

Some literature pays attention on CH selection. Reference [18] adopted a similar clustering CDG scheme as in [17]. The difference is that the method of selecting CH is improved, which introduces the entropy theory. They consider the CH selection as a decision problem based on multiple criteria. In addition to residual energy ($E(s_j)$) and distance to BS ($D(s_j)$), the intra-to-inter distance ratio ($DR(s_j)$) is included in CH selection criteria, which can judge the soundness of clustering methods. The entropy coefficient of each criterion is calculated and used to weight the criterion in P, which is the product of the three criteria. Multicriteria decision analysis methods such as the weighted product model are used to solve the decision problem, i.e., sensor nodes with the largest P will be selected as CHs. Reference [19] also paid attention to CH selection by considering residual energy, distance to the sink, node density, and network average energy. Besides, they make the first attempt to introduce a heterogeneous energy model in CDG-based clustering. Reference [20] used an improved LEACH protocol in CDG. The LEACH protocol is a classical clustering method in WSNs. They improve the LEACH protocol to match the sparse CDG. The main innovation is that a new criterion of CH selection, i.e., compression ratio, is introduced in the algorithm.

These improved methods of CH selection have a certain effect on load balance and energy efficiency in the network. But they have limited benefit on cutting down data transmissions.

Table 2: Clustering schemes of CDG.

Ref.	CDG scheme	CHs selection	Cluster formation	Main contribution	Year
[17]	Hybrid CDG	k-median	Shortest distance	Analyze the optimal number of clusters	2014
[18]	Hybrid CDG	Multicriteria decision problem	Shortest distance	CH selection considering residual energy, distance to BS, and the intra-to-inter distance ratio	2018
[19]	Hybrid CDG	Consider multiple parameters	Shortest distance	CH selection considering residual energy, distance to the sink, node density, and network average energy	2020
[20]	Sparse CDG	Random	Shortest distance	Improve the LEACH protocol to match the sparse CDG and CH selection considering compression ratio	2019
[21]	Plain CDG	Based on spatial location and density of node distribution	Shortest distance	Pay attention to load balance and design even clustering algorithms	2018
[22]	Plain CDG	Optimal solution based on the analytical model	Shortest distance	An analytical model for load balance and optimal number of CHs; the concept of backup CH to reduce the energy consumption caused by rotation of CHs	2019
[23]	Sparse CDG	At the center of evenly divided areas	Shortest distance	Devote to address the hot-spot area problem and propose an annular routing backbone tree	2019
[24]	Sparse CDG	Spatial location	Shortest distance	Propose a hexagonal topology and exploit spatial correlation between sensors' data; sleep scheduling	2019
[25]	Plain CDG	Location of event source	Shortest distance	Propose a dynamic clustering algorithm considering characteristics of CS theory	2017
[26]	Hybrid CDG	Center of clusters	Based on compressibility	Clustering based on compressive ratios of sensing data, greatly reduce data transmissions	2017
[27–31]	Sparsest and sparse CDG	Not mentioned	Not mentioned	Focus on designing measurement matrices to cut down data transmissions, resist packet loss, or ease implementation in hardware	2014-2020
[32–34]	Sparse CDG	Not mentioned	Not mentioned	Focus on exploiting spatiotemporal correlation by clustering to cut down data transmissions	2014-2019
[32]	Plain CDG	Not mentioned	Not mentioned	Propose a cluster size load balance technique for optimal utilization of CDG	2019

A number of literature focused on the load balance problem. In traditional clustering methods, most of CH selection are random such as LEACH protocol or based on some competition criteria such as in [18, 19]. However, these methods cannot guarantee even distribution of CHs, which can lead to better load balance of networks. Reference [21] paid attention to even distribution of CHs in clustering CDG. Two even clustering algorithms based on spatial location (LEC) and node distribution density (DEC) are proposed. The sensing area is divided into several equal grids. LEC is adopted in WSNs with uniform node distribution. In each grid, a CH is selected based on node's remaining energy and the distance from the node to the sink. DEC is adopted in WSNs with nonuniform node distribution. Node distribution density is measured by the grid. If density of a grid is lower than the threshold, the grid is merged into its neighboring grids whose density is higher than the threshold. Then, a CH is selected in each grid similar as in LEC. In clustering stage, CMs choose the closest CH and form clusters.

Reference [22] also focused on load balance in clustering CDG. The difference is that a sector-shaped network model is used which can be an absolute monitor area or just a part of larger general region. Therefore, the model can be applied to the network of any shape. The sector area is divided into k layer of ring-shaped area. To make sure load balance, it is supposed that the energy consumption of each layer is approximately equal. Under this assumption, the optimal number of CHs and optimal distribution of CHs in each layer are deduced with the constraint of minimizing the energy consumption in each layer. Besides, the concept of backup CH (BCH) was introduced in [22] to reduce the energy consumption caused by rotation of CHs. During the process of CH selection, the node with the second largest value of ratio of energy to distance (the criterion of CH selection) is preserved as BCH by each present CH. The rotation of CH happened between the current CH and BCH. The intracluster data were collected through the plain CDG method

The near sink area in WSNs is called the hot-spot area. In previous routing and clustering methods, the amount of data transmissions in the hot-spot area is much larger, causing load unbalance. The proposed CDG technologies relieve the problem to some extent, but the problem still exists. Reference [23] devoted to address the hot-spot area problem to achieve load balance. They proposed an annular routing backbone tree. Evenly distributed nodes in WSNs are equally divided into a number of clusters, except to those nodes within one hop from the sink, which send data directly to the sink. The CH is located at the center of each cluster. The sparse CDG is applied both in intracluster and intercluster data transmissions. Taking the sink as center, CHs with the same hops to the sink forms a number of rings.

The CH farthest from the sink is specified as the initial data collection node in each ring. From the initial node, data is gathered along the ring according right-hand rule until back to the initial node. Then, the initial node in the outmost ring sends data to the inner layer and finally to the sink through the shortest path. The initial nodes are selected again every r rounds, and clusters are reclustered every h rounds. In this way, within the range of one hop to the sink, only one node bears data packets of M measurements. The average data capacity of the near-sink node is significantly cut down; thus, the network's lifetime is extended.

Nevertheless, when designing clustering schemes, the above methods only consider characteristics of networks but neglect traits of CDG.

One of the advantages of clustering is that we can group nodes into different clusters for a certain purpose. That means we can exploit spatial correlation of sensors' data to reduce data transmission, since those correlated data is usually sparser in some transform domain. According to CS theory, those correlated data can be recovered by less CS measurements. So when designing a clustering algorithm, we can take advantage of data characteristics.

Reference [24] proposed a novel clustering CDG method which exploited spatial correlation between sensors' data to decrease data transmissions. Firstly, they form a hexagonal topology that is similar in cellular networks, which is beneficial to the scalability of networks. CHs are located close to intersections of the distance r_t to the BS and 60° angle lines of the BS, in which r_t is the transmission range of sensor nodes and the base station (BS). The selected CHs can expand the network by continuing to choose CHs at their r_t distance through the similar way. The radius of a cluster is equal to $r_t/2$, because sensory data of those nodes that are geographically close to each other is more spatially correlated. That means less required CS measurements need to be transmitted intraclusters. So in [24], only the CH senses and compresses data and then transmits data to the BS, which saves the intracluster communication cost. Other nodes in the cluster are sleep. However, in some scenarios, data from near nodes is not strongly correlated.

Most of WSNs are event-driven networks deployed to monitor fire disaster, earthquake, landslide, and so on. These abnormal events are explosive and usually burst at some locations, which can be referred to as event source. The event source will impact the local spatial correlation of sensor readings. Zhang et al. modeled the impact of event source on spatial correlation through JSM-1 and proposed a compressed sensing-based dynamic clustering algorithm centered on event source (CS-DCES) for CDG in [25]. The sink calculates the location of event source. Nodes around event source form a cluster. Because the intracluster data is more relevant, the sink recovers original data within a cluster in the CS-DCES algorithm, which requires less CS measurements. Simulation results indicate that under the same recovery accuracy, there is lower cost of data transmissions and more balanced load in CS-DCES. Reference [25] is one of few references that designed the clustering CDG algorithm considering the characteristic of CS theory. However, in reality, abnormal events are occasional and minority. In most cases, there is no abnormal event in the network. In this situation, the proposed algorithm will waste computational resources.

Lan and Wei proposed a compressibility-based (CBCA) clustering algorithm in [26]. The sensor nodes are converted to logic chains. Compressive ratios (CRs) of sensing data within a sliding window are circulated. Then, the network is clustered according to the CRs. If the CR of a cluster is less than the threshold, CDG is adopted in this cluster. Otherwise, the raw data gathering (RDG) is used. Simulation confirmed that the data transmission in CBCA is less compared to the random clustering method. However, CBCA supposes that all networks can be modeled as logic chains. In fact, only chain type or mesh network meet the requirement, which limits the application range of CBCA.

Wu et al. studied data characteristics and proposed the sparsest random scheduling for the CDG scheme in [27]. And Sun et al. proposed the sparsest clustering CDG (SRS-CCDG) in [28]. In [27], in order to improve the data sparsity, the abnormal sensing data, which is out of the range of predefined thresholds, is removed and sent directly to the sink. The sparsest CDG algorithm handles the smooth normal data (within the range of thresholds). They designed the sparsest measurement matrix $\mathbf{\Phi_e}$ with only one nonzero element in each row. And then, the matching orthogonal represent basis $\mathbf{\Psi_G}$ is constructed according to the RIP rule. So each measurement needs only one node to participate, which greatly cuts down data transmissions in the network. Clustering methods are good at partitioning the sensing data. So based on the work in [27], Reference [28] integrated the sparsest CDG with clustering to reduce data transmissions. Reference [29] exploited the sparsest CDG to resist packet loss in lossy WSNs. At the meantime, the sink constructs a block diagonal matrix (BDM) with the elements in the principal diagonal which are submeasurement matrices of each cluster. The sink reconstructs the entire network data using the BDM, which can take advantage of spatial correlation between clusters for the better recovering accuracy. However, the accuracy of recovery data in the sparsest CDG is not as precise as in dense projection such as plain CDG in the same framework. So it can be adopted in scenarios that do not require high data recovery accuracy. Besides, Reference [29] is the earlier research paying attention on the packet loss problem in CDG.

Reference [30] introduced a sparse circulant matrix as the measurement matrix in intracluster data transmission. The circulant matrix is structural which is easy to implement in hardware and saves the storage space of nodes. Reference [31] collected CS measurements through random sampling intracluster and random walk among CHs to reduce data transmissions in clustering CDG. They proved that the constructed measurement matrix was the adjacency matrix of an unbalanced expander graph which owned better data reconstruction accuracy. Both methods in [27–31] cut down network data transmissions and promote energy efficiency by reducing participated nodes in CDG

The same measurement matrix in [31] was used to exploit the spatiotemporal correlation of sensing data in CDG in [33]. The difference is that the random sampling

was adopted to choose the sensory data in the temporal domain. Considering both spatial and temporal correlation of sensing data can enhance accuracy of recovery data and reduce required CS measurement M. So Reference [34] combined Kronecker compressed sensing (KCS) and cluster topology to exploit spatial and temporal correlations simultaneously. The cluster topology can exploit spatial correlation of sensing data. And the collecting data of each node is a data vector in multiple continuous time slots. It is ensured that the collected data are temporally correlated. So it is a 2-dimensional compressive sensing problem and can be solved by KCS. Besides, the BDM in [29] was used in [34] to take advantage of the spatial correlation among clusters.

Reference [35] noticed that sensory data was always unordered and traditional sparsification bases such as DCT were inefficient to deal with unordered data. They introduced a Treelet transform into the clustering CDG. Treelet transform is used for multiscale analysis to unordered data. It can return an orthogonal basis reflected the internal localized correlation structure of the data. So the network is clustered according to the result of Treelet transform. Sensing data of nodes in the same cluster has strong correlation and can be recovered individually by less required CS measurements M. However, the proposed scheme needs training process which will cause delay in CDG. And there is an assumption that similar data of those nodes far away from each other is abnormal, which is not always established

Reference [32] considered that in some conventional clustering CDG schemes if the size of some clusters is too small, it is inefficient to implement CDG. So they proposed a cluster size load balance technique for optimal utilization of CDG by keeping the minimum number of nodes in clusters at level M. Though it is a small adjustment, the method means deep combination of clustering and CDG. Besides, a chicken swarm optimization algorithm is used to optimize the CS matrix to improve the reconstruction process, which means some metaheuristic algorithms are introduced in CDG and we will mention them in the next section

5. CDG Combined with Other Technologies

With the progress of researches in CDG, besides routing design and cluster topology, a lot of other technologies are introduced into CDG for further performance improvements. For example, the mobile data collection, distributed data storage (DDS), and energy harvesting are proved to be effective for energy efficiency and robustness in WSNs for IoTs. Many metaheuristic algorithms and learning methods are good at searching optimal or dynamic solution. We present an overview of CDG combined with other technologies in Table 3.

Mobile data gathering uses one or more mobile collectors to acquire sensing data, which can reduce energy expenditure of nodes in WSNs. The reason is that a mobile collector can walk closely to a node to collect data, which can save communication energy consumption of the node. Furthermore, in mobile data gathering, nodes only need to communicate with a mobile collector. So it is unnecessary to maintain the connectivity of nodes in an entire network. Due to its merits, there are several references introducing mobile collectors to the CDG. References [36, 37] used a mobile collector to acquire CS measurements from sensor nodes in a random walk method. In [36], an improved Metropolis-Hastings algorithm was adopted as a RW mobility scheme of the mobile collector for more uniform sampling distribution. The collected data from nodes visited by the mobile sink is temporal CS measurements. So the proposed method also exploits spatial-temporal correlation of collected data that is similar with [34]. Reference [37] adopted the standard random walk strategy and the kernel method to develop a sparse representation basis. Reference [38] used a mobile sink in clustering CDG. The mobile sink traverses the network to collect CS measurements from CHs. The hybrid CDG is applied intraclusters. The optimal traversing paths with the minimum length and energy consumption are calculated.

Traditional mobile collectors are always on the ground, which may be obstructed in the moving process. Unmanned aerial vehicles (UAVs) have attracted increasing attention in the telecom industry recently. Some scholars take UAVs as mobile collectors. Compared to traditional mobile collectors, UAVs have better flexibility to traverse the network to collect sensing data. Moreover, UAVs can fly to hard-to-reach areas.

Ebrahimi et al. used a UAV to gather data from CHs in clustering CDG in [39]. Nodes in the sensing area are clustered firstly. Sparse CDG acts as the data gathering scheme. CMs send their sparse projections of intracluster data along forwarding trees to CHs. The gathered data at the CHs are collected by the UAV, and then, the UAV deliver collected data to the sink. The UAV trajectory is well designed ensuring that the UAV can traverse all the CHs through the shortest path. The joint problem of optimizing node clustering, forwarding tree construction, CH selection per cluster, and UAV trajectory planning was studied. Lin et al. also adopted a UAV to collect CS measurements among CHs in [40]. They used hybrid CDG intraclusters. The UAV path planning optimization problem with the shortest traversing route is treated as a classical travelling salesman problem which is solved by the ACO algorithm.

In addition to the mobile sink, Zhang et al. introduced distributed data storage (DDS) technology to resist the impact of packet loss in CDG in [41]. DDS is introduced to enable reliable data gathering by employing redundancy. And CDG can compensate for the large amount of data transmissions caused by DDS to ensure sufficient redundancy. In the proposed scheme, some nodes broadcast their data packets to their neighbors with a certain probability. The neighbor nodes receive broadcasting data packets and add the packets to their own data packet. After repeating a few times, the DDS is completed and a mobile sink randomly queries M nodes and collects their stored data as CS measurements. The collected data is DDS data which can avoid failed nodes. But the DDS is implemented in probability. It is possible that there are some nodes without DDS data. If the mobile sink visits these nodes, the data recovery accuracy will be declined.

Table 3: Routing protocols of CDG.

Ref.	CDG scheme	Combined techniques	Main contribution	Year
[36–38]	Sparse CDG and hybrid CDG	Mobile collectors and mobile sink	Mobile collectors and mobile sink can save nodes' energy consumption of transmission	2017-2020
[39, 40]	Sparse CDG and hybrid CDG	UAV	Take UAV as a mobile data collector which has better flexibility to traverse the whole coverage area of WSN	2019-2021
[41]	Sparse CDG	DDS and mobile sink	DDS is introduced to enable reliable data gathering by employing redundancy	2018
[42–44]	Sparsest CDG and sparse CDG	Energy harvesting and sleep scheduling	Putting those nodes that do not take part in CS measurement into status of energy harvesting or sleep scheduling will prolong lifetime of WSN	2019-2020
[45]	Plain CDG	Grey Wolf optimization (GWO)	A metaheuristic algorithm GWO is used to search the optimal backbone tree connecting CHs to the sink and the optimal sensing matrix	2020
[46]	Sparse CDG	Bees algorithm and genetic algorithm (GA)	Incorporate Bees algorithm with genetic algorithm to search the optimal CS reconstructed data	2021
[47]	Sparse CDG	Multiple objective GA	Use multiple objective GA to calculate the optimal number of CS measurements and transmission range and sensing matrix	2021
[48, 49]	Sparse CDG	Dictionary learning, training algorithms	Use dictionary learning and training algorithms to get the optimal sparse basis	2020
[50]	Sparse CDG	Seed estimation algorithm	Estimate an adaptive seed used to generate the best random measurement matrix with minimal reconstruction error	2020
[51]	Sparse CDG	Reinforcement leaning (RL)	Use FRS-RL algorithm to select data aggregator nodes	2020
[52–55]	Sparse CDG	Deep learning (DL) and multiagent RL	Train deep neural networks to obtain a learned measurement matrix, reconstruction data with high accuracy, and the optimal compression ratio	2019-2021
[56–58]	Sparse CDG	DL	Use DL to extract information from compressed data and analyze compressed signal	2017-2021
[59, 60]	Sparse CDG	Edge computing	Edge computing helps with data secure algorithms and learning algorithms in CDG	2020-2021

Energy harvesting and sleep scheduling are also commonly used technologies in WSNs for energy saving. So some references attempt to introduce them in CDG for further performance improvement. Reference [42] adopted a partial canonical identity (PCI) measurement matrix that is similar to the measurement matrix in sparest CDG in [27]. Only a few nodes participate in the data gathering process, and the rest nodes harvest the energy. Reference [43, 44] adopted sleep scheduling in CDG. Reference [43] adopted a clustering CDG scheme and performs sleep scheduling for cluster member nodes. Considering two closely located nodes within the predetermined threshold distance, one of the nodes is set sleeping until the other node drains out of energy. Reference [44] proposed a fully distributed sleep scheduling method in a RW-based CDG which is named "sleeping CDG." All nodes in the network are active with the probability of p^{active}. It is ensured that in each round of data gathering, all the nodes are active once and only once, which can assure the successful recovery of data. In sparse CDG, only a few nodes participate in collecting CS measurements and the rest nodes can conduct energy harvesting or sleep for energy efficiency of the network. So combining sparse CDG with energy harvesting and sleep scheduling is viable.

Intelligent metaheuristic algorithms are good at some optimization problems. So there are a lot of researches aiming at solving optimization problems in CDG, such as optimizing routing paths, the measurement matrix, and recovering data, by these heuristic algorithms. Aziz et al. introduced a Grey Wolf optimization (GWO) algorithm to search the best path for each CH to the sink with minimum energy consumption in clustering CDG in [45]. The GWO algorithm simulates the Grey Wolf's nature in terms of hierarchical leadership and hunting behavior. Positions of CHs in the network are defined as positions in the hunting process, and the BS is defined as the position of the prey. The best paths with the least total distances and hops from each CH to the BS are searched by the GWO algorithm. The GWO algorithm is also used in searching the optimal sensing matrix with minimal mean square error of recovering data.

Traditional greedy algorithms are usually adopted in CS reconstruction problem. But it is hard to achieve the optimal solution. The metaheuristic-based reconstruction approaches are found to be potent in providing efficient solution. For example, the Bees algorithm was adopted in [18] to find optimal CS recovering data. And Reference [46] improved the Bees algorithm and also used it to determine optimal results for CS recovery in clustering CDG. They intended to incorporate the merits of the Bees algorithm and genetic algorithm (GA) by introducing crossover and mutation operators of the

genetic algorithm to the Bees algorithm. The metaheuristic algorithms definitely enhance recovering accuracy but increase computational complexity. Maybe the advance of hardware could compensate for the increased computational complexity to some extent. However, the proposed schemes are preliminary works that do not go deep into the incorporation of the two.

In addition, the metaheuristic algorithm can not only be used for a certain phase in the CDG process but also be optimized during the whole process. Reference [47] used a multiple objective genetic algorithm (MOGA) to calculate the optimal number M_{opt} of CS measurements, transmission range R_{opt}, and sensing matrix $\mathbf{\Phi}_{\text{opt}}$ at the BS in CDG. Then, the BS constructs M_{opt} paths and broadcasts $\mathbf{\Phi}_{\text{opt}}$ over the network.

Various learning algorithms have gained extensive attention recently. There are also a lot of references introducing learning methods in CDG. The learning-based CDG methods can be more dynamic and adaptive. The parameters in CDG can be modified with environment change. Reference [48] used the dictionary learning method to obtain a sparse basis in CDG. Most of previous CDG methods always adopt fixed sparse bases which sometimes are not suitable for diverse scenarios in WSNs. The proposed method learns from the training data to obtain the sparse basis which has better sparse representation ability. To address the overfitting of training data in the process of dictionary learning, the self-coherence penalty term is introduced in the algorithm. As a result, the learned sparse basis has low self-coherence structure, which can effectively suppress the impact of environmental noise. However, the training data in [48] is uncompressed historical data that needs a large number of communication resources and lacks of dynamic adaptability in time. So Reference [49] presented a training method to learn the sparse basis from CS measurements. In this way, the sparse basis can be updated in time and does not require extra data transmissions compared to [48]. Reference [50] got an adaptive dynamic random seed by a seed estimation algorithm. The adaptive seed is used to generate the best random measurement matrix which can result in the minimal reconstruction error. However, the seed estimation is conducted among those nodes near the sink whose results may not be the optimal seeds for the entire network.

Reinforcement learning (RL) and deep learning (DL) are hot technologies in Artificial Intelligence (AI) field at present. Introducing RL and DL into CDG can make data gathering intelligent and adaptable. By far, a few references extend research on using RL and DL in CDG. Reference [51] used a fuzzy rule system-based RL (FRS-RL) algorithm to select data aggregator nodes, which are used to assist CHs sending data to the mobile sink. The entire network is taken as an environment. CMs are learning agents, and choosing a root node is the action. The current node's link quality is the state. The Q-learning algorithm is used in the FRS-RL, and those CMs with the nearer distance and better link quality to the CHs likely become the data aggregator nodes. However, the RL algorithm in [51] has small contribution in the process of CDG. There is another metaheuristic algorithm used to find optimal paths for the mobile sink, which leads to high computational complexity.

References [52, 53] incorporated DL into CDG and designed a deep compressed sensing network (DCSNet) to build a measurement matrix and reconstruct data from CS measurements. The authors use an end-to-end learning method to train a deep neural network. There are two mappings that need learning. The first mapping is from the raw data in sensor domain to the compressed data in latent domain and will obtain a learned measurement matrix. Using the learned measurement matrix the abstract features of data are conserved in the CS measurements, which can achieve higher accuracy of recovery data. The second mapping is from the compressed data in latent domain to the reconstructed data and obtains a high-performance reconstructor. That means the main processes of CDG have applied the DL algorithm, which make the CDG scheme more intelligent. Considering the implementation of DCSNet in resource-limited WSNs for IoTs, the training is offline

Reference [54] noticed that data sparsity is variable in most scenarios due to its time-varying nature, so they changed the compression ratio of CDG in real time with data sparsity to transmit data more efficiently. They proposed a lightweight and adaptive compressed sensing method based on deep learning for edge devices (LACSLE) using a pretrained deep learning model for mapping sensor data to an optimal compression ratio. A supervised learning and a reinforcement learning method are used in the proposed scheme, respectively, which are denoted as LACSLE-SL and LACSLE-RL. LACSLE-RL does not need training data but the accuracy of estimating an optimal compression ratio is lower than that of LACSLE-SL. However, LACSLE is adopted in a single sensor (edge device) and only exploits temporal correlation of sensor data. So they extended their work and proposed LADICS-MARL for multiple sensors in [55]. Multiagent deep reinforcement learning is used. The optimal compression ratio is based on data from multiple sensors, and an optimal combination of compressed data is recovered simultaneously. As a result, LADICS-MARL can achieve higher data recovery accuracy and transmission efficiency compared to LACSLE

In addition to data recovery and some optimal problems in CDG, machine learning can also be used to analyze data. Reference [56] applied compressive learning, which used machine leaning algorithms to extract information from compressed signal, to spectrum sensing for cognitive radios. The channel occupancy information is collected by sensors, compressed, and transmitted to the fusion center. A detector based on neural networks is used to extract channel occupancy information. Compared to the optimum maximum likelihood detector, the proposed scheme has lower complexity. Reference [57] proposed a deep learning classification which can directly classify biomedical EEG signal from compressive measurements. This is an integrated approach for signal reconstruction and analysis in a wireless body area network. They incorporated a classifier into the deep BCS, which is an unsupervised feature extraction tool in [58].

The above works show that the learning algorithms have great potential to provide a global and intelligent solution in CDG, which is worth doing in-depth researches. This is a fascinating research point. However, learning algorithms also consume a large amount of computational resources, which is a big challenge for resource-limited WSNs. Edge computing may be a solution. In fact, there are already some works devoted to introduce edge computing in CDG, such as in [54, 55]; the process of learning is performed in edge servers. Reference [59] proposed a secure CS data transmission framework with the help of edge computing. The strong encryption algorithm is transferred to the edge cloud. Reference [60] proposed a novel edge computing framework FCL in PM2.5 air quality monitoring systems on large-scale smart city sensing application. FCL integrates federated learning into CDG, which can reduce the network congestion while maintaining data privacy. An edge computing framework is designed to assist in implementing federated learning. To sum up, the emerging edge computing technique could help to further improve the performance of CDG.

6. Conclusion

As the perceptual layer of IoTs, WSNs are responsible for collecting sensing data for upper layers. An efficient data gathering method is crucial for WSNs. CDG offers an optimal scheme. In this paper, we summarize the published CDG methods in WSNs from three aspects, i.e., routing protocol of CDG, clustering scheme of CDG, and CDG combined with other technologies. All these algorithms aim at one or multiple objectives such as reducing data transmission, balancing load traffic, improving data recovery accuracy, promoting energy efficiency and dynamic adaptability, and prolonging lifetime of WSNs for IoTs. We analyze merits, disadvantages, and limitations of each kind of algorithm.

The arrangement of references is basically in chronological order, from which we can trace the developing process of CDG and predict future research directions. The early work mainly commits to realize CDG in applicable manners. Three types of basic framework named plain CDG, hybrid CDG, and sparse CDG are formed. And two categories of routing protocols in CDG, i.e., determined routing and RW routing, are designed in energy-efficient methods. Determined routing is controllable and can be optimized with some objective functions. But it is more computationally intensive and lacks of dynamic adaptability, while RW routing acts on the contrary and is more suitable for sparse CDG. Then, in the next stage, clustering topology control technologies are introduced in CDG. The clustering CDG is good at balancing traffic load and grouping nodes into clusters for reduction of required CS measurements by exploiting spatial and temporal correlation of sensing data. At last, with progresses of technology, various methods are integrated in CDG aiming at establishing green and intelligent WSNs for IoTs.

However, it is a pity that the combination of various algorithms in CDG is simple and fundamental. That is mainly reflected in the following aspects. For one thing, some complex algorithms, such as metaheuristic algorithms and RL algorithms, are adopted merely in one or two procedures of CDG schemes, which have limited performance enhancement to CDG but consume many computational resources. For another, according to respective characteristics and internal relationships of CDG and other algorithms, joint and global optimization needs to be studied. It is also the development trend of CDG in the future.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (62061009 and 62161031), the Guangxi Natural Science Foundation (2020GXNSFBA297097), the Fund of Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing (GXKL06200102), and the Innovation Project of Guangxi Graduate Education (YCBZ2020061).

References

[1] K. Kaur, "A survey on Internet of things-architecture, applications, and future trends," in *2018 First International Conference on Secure Cyber Computing and Communicaion (ICSCCC)*, pp. 581–583, Jalandhar, India, 2018.

[2] C. Luo, F. Wu, J. Sun, and W. C. Chen, "Compressive data gathering for large-scale wireless sensor networks," in *Fifth ACM International Conference on Mobile Computing and Networking (MOBICOM)*, pp. 145–156, Beijing China, 2009.

[3] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[4] E. J. Candès, "La propriete d'isometrie restreinte et ses consequences pour le compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[5] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[6] C. Luo, F. Wu, J. Sun, and W. C. Chen, "Efficient measurement generation and pervasive sparsity for compressive data gathering," *IEEE Transactions on Wireless Communications*, vol. 9, no. 12, pp. 3728–3738, 2010.

[7] J. Luo and L. Xiang, "Does compressed sensing improve the throughput of wireless sensor networks?," in *2010 IEEE International Conference on Communications*, Cape Town, South Africa, 2010.

[8] W. Wang, M. Garofalakis, and K. Ramchandran, "Distributed sparse random projections for refinable approximation," in *6th International Symposium on Information Processing Sensor Networks*, pp. 331–339, Cambridge Massachusetts USA, 2007.

[9] D. Ebrahimi and C. Assi, "Optimal and efficient algorithms for projection-based compressive data gathering," *IEEE Communications Letters*, vol. 17, no. 8, pp. 1572–1575, 2013.

[10] D. Ebrahimi and C. Assi, "A distributed method for compressive data gathering in wireless sensor networks," *IEEE Communications Letters*, vol. 18, no. 4, pp. 624–627, 2014.

[11] D. Ebrahimi and C. Assi, "Compressive data gathering using random projection for energy efficient wireless sensor networks," *Ad Hoc Networks*, vol. 16, pp. 105–119, 2014.

[12] S. Abbasi-Daresari and J. Abouei, "Toward cluster-based weighted compressive data aggregation in wireless sensor networks," *Ad Hoc Networks*, vol. 36, pp. 368–385, 2016.

[13] H. F. Zheng, F. Yang, X. H. Tian, X. Y. Gan, and X. B. Wang, "Data gathering with compressive sensing in wireless sensor networks: a random walk based approach," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 1, pp. 35–44, 2015.

[14] M. T. Nguyen and K. A. Teague, "Compressive sensing based random walk routing in wireless sensor networks," *Ad Hoc Networks*, vol. 54, pp. 99–110, 2017.

[15] J. J. Huang and B. H. Soong, "Cost-aware stochastic compressive data gathering for wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1525–1533, 2019.

[16] P. Zhang and J. X. Wang, "On enhancing network dynamic adaptability for compressive sensing in WSNs," *IEEE Transactions on Communications*, vol. 67, no. 12, pp. 8450–8459, 2019.

[17] R. Xie and X. Jia, "Transmission-efficient clustering method for wireless sensor networks using compressive sensing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, pp. 806–815, 2014.

[18] W. Osamy, A. M. Khedr, A. Aziz, and A. A. El-Sawy, "Cluster-tree routing based entropy scheme for data gathering in wireless sensor networks," *IEEE Access*, vol. 6, pp. 77372–77387, 2018.

[19] R. Manchanda and K. Sharma, "Energy efficient compression sensing-based clustering framework for IoT-based heterogeneous WSN," *Telecommunication Systems*, vol. 74, no. 3, pp. 311–330, 2020.

[20] Y. Song, Z. G. Liu, X. L. He, and H. Jiang, "Research on data fusion scheme for wireless sensor networks with combined improved LEACH and compressed sensing," *Sensors*, vol. 19, no. 21, p. 4704, 2019.

[21] J. H. Qiao and X. Y. Zhang, "Compressive data gathering based on even clustering for wireless sensor networks," *IEEE Access*, vol. 6, pp. 24391–24410, 2018.

[22] Q. Wang, D. Y. Lin, P. F. Yang, and Z. Q. Zhang, "An energy-efficient compressive sensing-based clustering routing protocol for WSNs," *IEEE Sensors Journal*, vol. 19, no. 10, pp. 3950–3960, 2019.

[23] Y. C. Yuan, W. Liu, T. Wang, Q. Y. Deng, A. F. Liu, and H. B. Song, "Compressive sensing-based clustering joint annular routing data gathering scheme for wireless sensor networks," *IEEE Access*, vol. 7, pp. 114639–114658, 2019.

[24] U. S. Pacharaney and R. K. Gupta, "Clustering and compressive data gathering in wireless sensor network," *Wireless Personal Communications*, vol. 109, no. 2, pp. 1311–1331, 2019.

[25] C. Zhang, X. Zhang, O. Li, Y. P. Yang, and G. Y. Liu, "Dynamic clustering and compressive data gathering algorithm for energy-efficient wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 13, no. 10, Article ID 155014771773890, 2017.

[26] K. C. Lan and M. Z. Wei, "A compressibility-based clustering algorithm for hierarchical compressive data gathering," *IEEE Sensors Journal*, vol. 17, no. 8, pp. 2550–2562, 2017.

[27] X. G. Wu, Y. Xiong, P. L. Yang, S. H. Wan, and W. C. Huang, "Sparsest random scheduling for compressive data gathering in wireless sensor networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5867–5877, 2014.

[28] P. Sun, L. T. Wu, Z. B. Wang, M. Xiao, and Z. Wang, "Sparsest random sampling for cluster-based compressive data gathering in wireless sensor networks," *IEEE Access*, vol. 6, pp. 36383–36394, 2018.

[29] C. Zhang, O. Li, Y. P. Yang, G. Y. Liu, and X. Tong, "Corrigendum to "Energy-efficient data gathering algorithm relying on compressive sensing in lossy WSNs" [Measurement 147 (2019) 106875]," *Measurement*, vol. 149, article 107099, 2020.

[30] N. Wang, D. Chen, J. Y. Chen, X. Xu, and J. W. Wan, "Clustering data gathering method based on compressed sensing in wireless sensor networks," in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 214–217, Hangzhou, China, 2018.

[31] X. L. Li, X. F. Tao, and G. Q. Mao, "Unbalanced expander based compressive data gathering in clustered wireless sensor networks," *IEEE Access*, vol. 5, pp. 7553–7566, 2017.

[32] A. Aziz, K. Singh, W. Osamy, and A. M. Khedr, "Effective algorithm for optimizing compressive sensing in IoT and periodic monitoring applications," *Journal of Network and Computer Applications*, vol. 126, pp. 12–28, 2019.

[33] X. L. Li, X. F. Tao, and Z. Chen, "Spatio-temporal compressive sensing-based data gathering in wireless sensor networks," *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 198–201, 2018.

[34] C. Zhang, O. Li, X. Tong, K. Ke, and M. X. Li, "Spatiotemporal data gathering based on compressive sensing in WSNs," *IEEE Wireless Communications Letters*, vol. 8, no. 4, pp. 1252–1255, 2019.

[35] C. Zhao, W. X. Zhang, X. M. Yang, and Y. Q. Song, "A novel compressive sensing based data aggregation scheme for wireless sensor networks," in *2014 IEEE International Conference on Comminications (ICC)*, pp. 18–23, Sydney, NSW, Australia, 2014.

[36] H. F. Zheng, J. Y. Li, X. X. Feng, W. Z. Guo, Z. H. Chen, and N. Xiong, "Spatial-temporal data collection with compressive sensing in mobile sensor networks," *Sensors*, vol. 17, no. 11, p. 2575, 2017.

[37] H. F. Zheng, W. Z. Guo, and N. X. Xiong, "A kernel-based compressive sensing approach for mobile data gathering in wireless sensor network systems," *IEEE Transactions on Systems Man Cybernetics-Systems*, vol. 48, no. 12, pp. 2315–2327, 2018.

[38] S. P. Tirani, A. Avokh, and S. Azar, "WDAT-OMS: a two-level scheme for efficient data gathering in mobile-sink wireless sensor networks using compressive sensing theory," *IET Communications*, vol. 14, no. 11, pp. 1827–1838, 2020.

[39] D. Ebrahimi, S. Sharafeddine, P. H. Ho, and C. Assi, "UAV-aided projection-based compressive data gathering in wireless sensor networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1893–1905, 2019.

[40] C. Lin, G. J. Han, X. Y. Qi, J. du, T. T. Xu, and M. Martinez-Garcia, "Energy-optimal data collection for unmanned aerial vehicle-aided industrial wireless sensor network-based agricultural monitoring system: a clustering compressed sampling approach," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4411–4420, 2021.

[41] C. Zhang, O. Li, G. Liu, and M. Li, "A practical data-gathering algorithm for lossy wireless sensor networks employing distributed data storage and compressive sensing," *Sensors*, vol. 18, no. 10, p. 3221, 2018.

[42] N. Jain, V. A. Bohara, and A. Gupta, "iDEG: integrated data and energy gathering framework for practical wireless sensor networks using compressive sensing," *IEEE Sensors Journal*, vol. 19, no. 3, pp. 1040–1051, 2019.

[43] R. Manchanda and K. Sharma, "A novel framework for energy-efficient compressive data gathering in heterogeneous wireless sensor network," *International Journal of Communication Systems*, vol. 34, no. 3, 2021.

[44] S. Mehrjoo, F. Khunjush, and A. Ghaedi, "Fully distributed sleeping compressive data gathering in wireless sensor networks," *IET Communications*, vol. 14, no. 5, pp. 830–837, 2020.

[45] A. Aziz, W. Osamy, A. M. Khedr, A. A. el-Sawy, and K. Singh, "Grey wolf based compressive sensing scheme for data gathering in IoT based heterogeneous WSNs," *Wireless Networks*, vol. 26, no. 5, pp. 3395–3418, 2020.

[46] A. Salim, W. Osamy, A. M. Khedr, A. Aziz, and M. Abdel-Mageed, "A secure data gathering scheme based on properties of primes and compressive sensing for IoT-based WSNs," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5553–5571, 2021.

[47] M. A. Mazaideh and J. Levendovszky, "A multi-hop routing algorithm for WSNs based on compressive sensing and multiple objective genetic algorithm," *Journal of Communications and Networks*, vol. 23, no. 2, pp. 138–147, 2021.

[48] J. Y. Chen, F. Q. Zhou, Z. S. Guo, and J. W. Wan, "Compressed data collection method for wireless sensor networks based on optimized dictionary updating learning," *IEEE Access*, vol. 8, pp. 205124–205135, 2020.

[49] P. Zhang, J. X. Wang, and W. J. Li, "A learning based joint compressive sensing for wireless sensing networks," *Computer Networks*, vol. 168, p. 107030, 2020.

[50] A. Aziz, K. Singh, W. Osamy, and A. M. Khedr, "An efficient compressive sensing routing scheme for internet of things based wireless sensor networks," *Wireless Personal Communications*, vol. 114, no. 3, pp. 1905–1925, 2020.

[51] G. Sanjay Gandhi, K. Vikas, V. Ratnam, and K. Suresh Babu, "Grid clustering and fuzzy reinforcement-learning based energy-efficient data aggregation scheme for distributed WSN," *IET Communications*, vol. 14, no. 16, pp. 2840–2848, 2020.

[52] M. Q. Zhang, H. X. Zhang, D. F. Yuan, and M. G. Zhang, "Compressive sensing and autoencoder based compressed data aggregation for green IoT networks," in *2019 IEEE global communications conference (GLOBECOM)*, Waikoloa, HI, USA, 2019.

[53] M. Q. Zhang, H. X. Zhang, D. F. Yuan, and M. G. Zhang, "Learning-based sparse data reconstruction for compressed data aggregation in IoT networks," *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11732–11742, 2021.

[54] M. Sekine and S. Ikada, "LACSLE: lightweight and adaptive compressed sensing based on deep learning for edge devices," in *2019 IEEE global communications conference (GLOBECOM)*, Waikoloa, HI, USA, 2019.

[55] M. Sekine and S. Ikada, "Adaptive cooperative distributed compressed sensing for edge devices: a multiagent deep reinforcement learning approach," in *19th IEEE International Conference on Pervasive Computing and Communications (IEEE PerCom)*, Kassel, Germany, 2021.

[56] P. H. C. de Souza, L. L. Mendes, and M. Chafii, "Compressive learning in communication systems: a neural network receiver for detecting compressed signals in OFDM systems," *IEEE Access*, vol. 9, pp. 122397–122411, 2021.

[57] V. Singhal, A. Majumdar, and R. K. Ward, "Semi-supervised deep blind compressed sensing for analysis and reconstruction of biomedical signals from compressive measurements," *IEEE Access*, vol. 6, pp. 545–553, 2018.

[58] S. Singh, V. Singhal, and A. Majumdar, "Deep blind compressed sensing," in *2017 Data Compression Conference (DCC)*, Snowbird, UT, USA, 2017.

[59] Y. Zhang, P. Wang, L. Fang, X. He, H. Han, and B. Chen, "Secure transmission of compressed sampling data using edge clouds," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6641–6651, 2020.

[60] K. T. Putra, H. Chen, M. R. Prayitno et al., "Federated compressed learning edge computing framework with ensuring data privacy for PM2.5 prediction in smart city sensing applications," *Sensors*, vol. 21, no. 13, p. 4586, 2021.