2021 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS) | 978-1-6654-1327-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/HPBDIS53214.2021.9658349

# A Survey on Capsule Networks: Evolution, Application, and Future Development

Jiawei Li[1], Qichen Zhao[2], Nan Li[3, 4]*, Lin Ma[3, 4], Xuan Xia[3, 4], Xiaoguang Zhang[3, 4], Ning Ding[3, 4], Nannan Li[5]
[1]*College of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, P. R. China.*
[2]*School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, P. R. China.*
[3]*Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, P. R. China.*
[4]*Institute of Robotics and Intelligent Manufacturing, The Chinese University of Hong Kong, Shenzhen, Shenzhen, P. R. China.*
[5]*International Institute of Next Generation Internet, Macau University of Science and Technology, Macau, P. R. China.*
*lijiawei5876@stu.xjtu.edu.cn , 119010459@link.cuhk.edu.cn,*
*{linan, malin, xiaxuan, zhangxiaoguang, dingning}@cuhk.edu.cn, nnli@must.edu.mo*

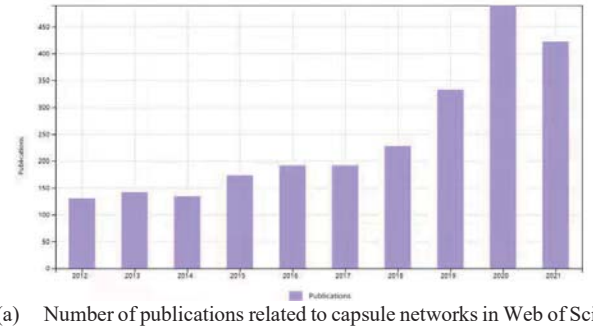
***Abstract* – Capsule networks have drawn increasing attention since the concept is proposed in 2011. They hold advantages over convolutional neural networks (CNNs) on modeling part-to-whole relationships between entities and learning viewpoint invariant representations. A variety of improved designs and applications related to capsule networks have been explored within the past decade. This paper presents a relatively thorough survey of literatures corresponding to the evolution of the capsule networks and their applications. Different design of capsule networks from Hinton's group as well as other researchers are reviewed. Applications of capsule networks in various scenarios and tasks are elaborated, including hyperspectral image processing, medical image processing, facial image processing, text classification, object detection, image segmentation, few-shot learning, etc. The relation and comparison of capsule networks, CNNs, and transformers, and the trend of future development of capsule networks are discussed. We believe this review provides a good reference to researchers who are interested in capsule networks.**

***Index Terms* –Capsule networks, deep learning, artificial intelligence, artificial neural network.**

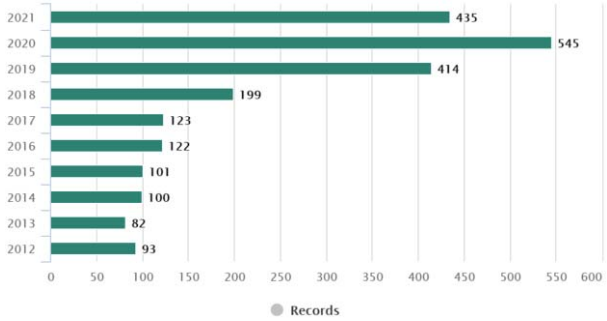
## I. INTRODUCTION

The ongoing upsurge of deep learning is mainly set off by the success of CNNs, which are the mainstream form of deep learning models, especially for computer vision tasks. However, CNNs are not designed to model the important spatial level correlation between different objects and structures. Besides, much of the spatial information is lost in the pooling operations in CNNs.

Hinton et al. [1, 2] propose the capsule networks to address the intrinsic limitations of CNNs. By mimicking the human visual system to obtain equivariance, instead of the original translational invariance, capsule networks can use less data to get more extensive generalization in different perspectives. Much attention has been drawn to the capsule networks in the deep learning research community since they provide a novel way of constructing deep neural network architectures. The exploration and application of capsule networks have shown a rapid upward trend since the concept is proposed, as depicted in Fig. 1. After nearly ten years of development, lots of new structural design and applications of capsule networks have been proposed.



(a) Number of publications related to capsule networks in Web of Science since 2012




(b) Number of publications related to capsule networks in Engineering Village since 2012

Fig. 1. Number of publications found in the Web of Science (a) and Engineering Village databases in recent 10 years.

This work is partially supported by National Key R&D Program of China (2021YFE0204200, 2019YFB1310403, 2019YFB1310402), National Natural Science Foundation of China (U1813216, U2013202, 61806190), foundation of Shenzhen Institute of Artificial Intelligence and Robotics for Society (AC01202101022), and the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2019A1515111119, 2021A1515010926, and 2021B1515420005.
*Corresponding author

978-1-6654-1327-5/21/$31.00 ©2021 IEEE

In this paper, we conduct a literature review to present the current status of the research related to capsule networks. We firstly introduce the evolution of capsule networks designed by Hinton's group, which is the leading group in this area. Novel designs and improvements of capsule networks devised by other researchers are introduced subsequently. Then, various applications of capsule networks are elaborated, including in the popular scenarios, such as hyperspectral image processing, medical image processing, facial image processing, text classification, electricity patrol, and other complex tasks like object detection, image segmentation and few shot learning. Finally, we discuss the relation and comparison of capsule networks, CNNs and transformers, as well as the future developmental trend of capsule networks. We believe this review provides a good reference to researchers who are interested in capsule networks.

## II. Capsule Networks and Variants

The core idea of capsule networks is to use capsules to perform complicated internal computations on the inputs, and then output vectors. Each capsule learns to recognize an implicitly defined visual entity over a limited domain of viewing conditions and deformations, and outputs both the probability that the entity is present and a set of "instantiation parameters" that may include the precise pose, lighting and deformation of the entity relative to an implicitly defined canonical version of that entity [1]. The implicitly defined canonical version of the entity can be regarded as a "standard entity". The combination of the standard entity and the instantiation parameter is the deformed entity caused by the change of viewpoint. When the capsule is working, the probability of the entity being present is locally invariant, and the instantiation parameters are "equivariant". As the viewpoint changes, the instantiation parameters will change.

In 2011, Hinton et al. [1] present the preliminary concept of capsule networks. Also, they construct a simple example of this method called Transforming Autoencoder, as shown in Fig. 2. A simple two-dimensional image translation is used as an example to verify the network. The capsules in the network only interact at the final layer. The network merges the outputs of these capsules to get the final result.

Sabour et al. [2] propose an architecture of a 3-layer capsule network in 2017, as shown in Fig. 3. The low-level capsules activate the high-level capsules through an iterative dynamic routing mechanism. They define a capsule as a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part. The length of the activity vector is used to represent the probability that the entity exists and its orientation to represent the instantiation parameters.

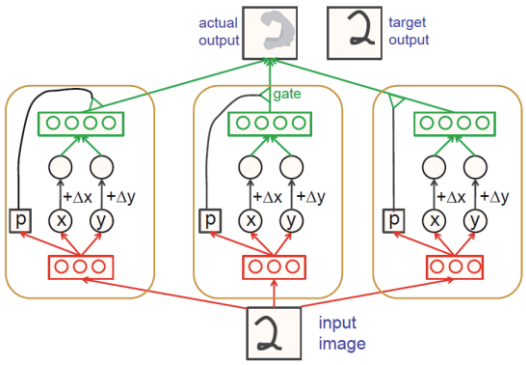



Fig. 2. Structure of a transforming auto-encoder with three capsules [1].

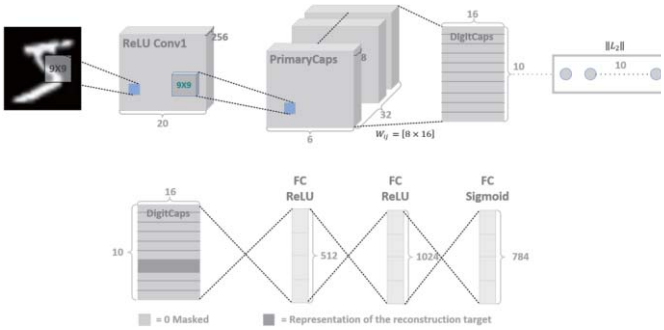



Fig. 3. Structure of the capsule network proposed by Sabour et al. in [2].

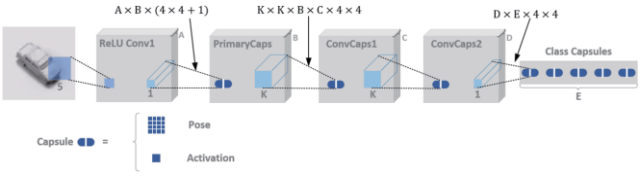



Fig. 4. Structure of the capsule network proposed by Hinton et al. in [3].

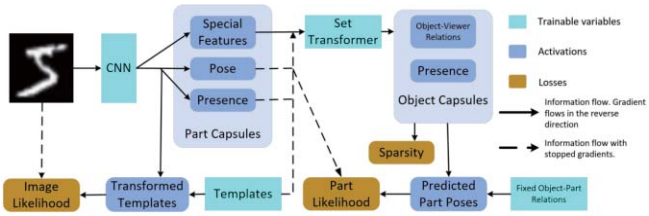



Fig. 5. Structure of the capsule network proposed by Kosiorek et al. in [4].

In 2018, Hinton et al. [3] introduce convolutional capsule layers into the capsule network and improve the routing process with the EM algorithm. They also improve the structure of the capsule, as shown in Fig. 4. Each capsule has a logistic unit for representing the presence of the entity, and a 4*4 matrix for representing the relationship between the entity and the observer.

Kosiorek et al. [4] make a change to the capsule network in 2019. They turn supervised learning into unsupervised learning, and introduced autoencoders, attention mechanisms into the capsule network. Moreover, the concepts of part capsules and object capsules are presented, as shown in Fig. 5. Its general idea is to extract the parts and their attributes from the image, then reconstruct the objects based on these parts and their attributes to discover the relationship between the parts and the objects.

Sun et al. [5] propose a self-supervised capsule architecture, the canonical capsules, for processing 3D point clouds in 2021. They compute capsule decompositions of objects through permutation-equivariant attention, and train the 3D deep representation in a truly unsupervised way. They achieve state-of-the-art performance in unsupervised 3D point cloud registration, reconstruction, and unsupervised classification.

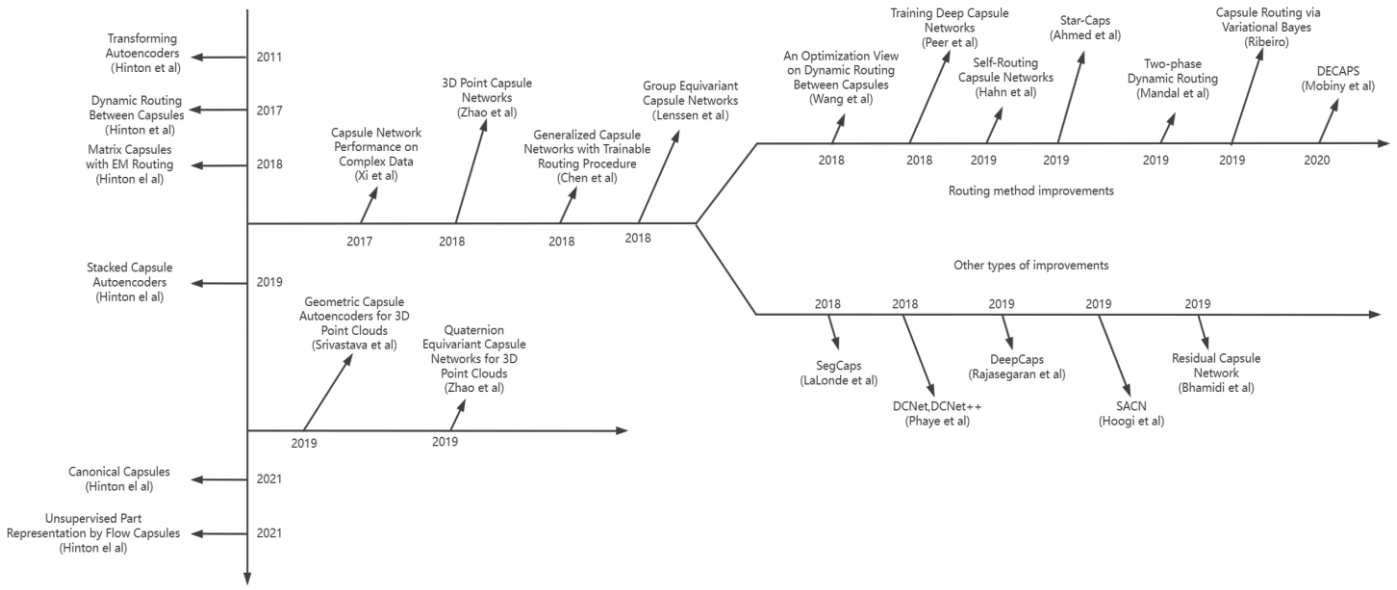



Fig. 6. Evolution of the structural design of capsule networks.

Sabour et al. [6] propose the FlowCapsules, a way to learn primary capsule encoders that detect atomic parts from a single image. The capsule encoder takes as input a single frame and estimates a set of primary capsules, each comprising a shape mask in canonical coordinates, a pose transformation from canonical to image coordinates, and a scalar representing relative depth. Learning is accomplished in an unsupervised manner, using flow estimation from capsule shapes and poses as a *proxy* task.

The above are the main evolution history of capsule networks designed by Hinton's group, which is the leading group of capsule network research. Besides, other researchers have also proposed their design and improvements on capsule networks.

Xi et al. [7] try to further test the application of the capsule network on high-dimensional data. They try various model modifications, including stacking more capsule layers, increasing the number of primary capsules, etc. The result showed that increasing convolution layers and using a 4-model ensemble will increase validation accuracy. Zhao et al. [8] use capsule networks to implement an auto-encoder to process sparse 3D point clouds. The capsule-encoder accepts point cloud as input. The decoder accepts the latent capsules output by the capsule-encoder as input, and reconstructs the point set according to the latent features. Chen et al. [9] find that the routing procedure of the capsule network is not well incorporated into the entire training process. Therefore, they add routing procedure to the training process. The general capsule networks do not have guaranteed equivariance or invariance. In order to solve this problem, Lenssen et al. [10] propose a new type of capsule called group equivariant capsule, and a scheme for dynamic routing algorithm.

One of the main directions to improve the capsule network is to improve the routing strategy. The iterative routing strategy of capsule networks is actually a clustering idea. The unsupervised nature of clustering will lead to high computational complexity. Besides, cluster assumption may not hold in presence of heavy input noise. Wang et al. [11] regard the routing strategy as an optimization process. This process needs to minimize the clustering-like loss and a KL regularization term between the current coupling distribution and its last states. Peer et al. [12] claim that the routing-by-agreement algorithm can not guarantee the presence of parse trees in the network. So they propose a routing algorithm called dynamic deep routing. Hahn et al. [13] design a new routing strategy based on Mixture-of Experts. Ahmed et al. [14] propose straight-through attentive routing to improve the routing mechanism of capsule networks. Mandal et al. [15] find that the micro and macro-level features have the same priority in the routing process. So they propose a two-phase dynamic routing protocol based on the hierarchical learning paradigm. Ribeiro et al. [16] propose a routing algorithm derived from Variational Bayes. Mobiny et al. [17] use an inverted dynamic routing mechanism to enable the capsules to selectively focus on small but informative details in the data. They also propose a training procedure called Peekaboo.

In addition to improving the iterative routing strategy of capsule networks, there are other ways to improve the capsule network. LaLonde and Bagci [18] solve the computationally expensive problems by extending the idea of convolutional capsules. They also expand the use of capsule networks to object segmentation tasks for the first time. Phaye et al. [19] replace the standard convolutional layers with densely connected convolutions to improve the feature extraction capability of capsule networks. Besides, they try to learn complex data efficiently by representing spatial information in a fine-to-coarser manner. Rajasegaran et al. [20] improve the dynamic routing algorithm, but their purpose is to increase the depth of the capsule network. Hoogi et al. [21] introduce the self-attention mechanism as an integral layer into the shallow capsule network for the first time to compensate for the lack of deep network. Bhamidi et al. [22] combine the residual network with

the capsule network. They replace the convolutional layer in the traditional capsule network with skip connections.

There are also other ways of modify the capsule networks. Srivastava et al. [23] propose a novel voting mechanism. They use auxiliary viewpoints to discover the value coded representation of the parent object. Zhao et al. [24] propose a 3D capsule module for processing point clouds. The network can disentangle geometry from pose. Besides, they establish a theoretical connection between the routing procedure and the Weiszfeld algorithm. They also propose a new routing algorithm based on this.

We summarize the evolution of the capsule networks both designed by Hinton's group and other researchers and illustrate the development in Fig. 6.

## III. Applications of Capsule Networks

Capsule networks have been frequently applied in many machine learning tasks ever since its appearance, for processing hyperspectral images, time sequence data, medical images, electricity patrol, texts, etc. This section summarizes various applications of capsule networks as shown in Table I.

### *A. Capsule Networks Used in Hyperspectral Image Processing*

Hyperspectral images often have high dimensions; thus very complex structure is required if CNNs are selected for classification. However, spatial and spectral information of hyperspectral images can be well captured by capsule networks. Therefore, capsule networks are frequently utilized in hyperspectral image classification. Zhang et al. [25] develop a 1D-convolutional capsule network which extracts spatial and spectral information separately, and reduces the parameter amount and computational complexity of conventional capsule networks. Wang et al. [26] explore the 1D-structure TripleGAN for sample generation and integrate capsule networks for hyper spectral image classification to propose a Caps-TripleGAN framework. Arun et al. [27] propose a framework combining Capsule Networks and LSTM for extracting spectral and spatial features to improve hyperspectral image classification. Deng et al. [28] present a modified two-layer CapsNet with limited training samples for hyperspectral image classification. Yin et al. [29] tune a new CapsNet architecture with 3 convolutional layers for hyperspectral image classification. A shallow layer is added to provide higher level features to the primary capsules. Majority of the attempts which try to adopt capsule networks into operating hyperspectral images have appeared to be successful, therefore proved the capability of capsule networks in this area.

### *B. Capsule Networks Used in Time Sequence Data Processing*

Capsule networks also show stunning performance when dealing with time sequence data, e.g., traffic flows, brain wave and seismic waves. Though 1D CNNs and RNNs have shown promising performance in understanding the sequence data, they may lose important information which capsule networks can capture and provide accurate prediction. Kim et al. [30] propose a neural network with capsules for traffic flow prediction in complex road networks. Yao et al. [31] use capsule networks in an end-to-end IoT traffic classification method that integrates feature extraction, feature selection, and classification model. Chao et al. [32] construct a deep learning framework based on a multiband feature matrix (MFM) and a capsule network. Frequency domain, spatial characteristics, and frequency band characteristics of the multi-channel EEG signals are combined to construct the MFM in the framework, then the CapsNet model is introduced for pattern recognition. Peng et al. [33] present an automatic microseismic record classifier with limited samples in underground mines based on CapsNet. The data is converted into feature matrix with respect to frame and commonly used feature in time and frequency domain. Wang et al. [34] is inspired by the brain anatomical structure and proposed a multi-kernel capsule network (MKCapsnet). Many of the works mentioned above in processing time sequence data compared capsule networks with CNNs. Results indicate that capsule networks outperform CNNs in this area most of the time, once again provide a credible evidence for capsule networks' ability.

**TABLE I. APPLICATIONS OF CAPSULE NETWORK**

| Application | Article | Structure | Preprocessing | Dataset |
|---|---|---|---|---|
| Hyperspectral image | [25] | Conv + PrimaryCaps + 1D-ConvCaps + ClassCaps | PCA-Whitening | Indian Pines, University of Pavia & Salinas |
| | [26] | TripleGAN + Relu Conv + PrimaryCaps + CategoryCaps | Unmentioned | Indian Pines AVIRIS, Pavia University ROSIS & Xuzhou HYSPEX |
| | [27] | Conv + PrimaryCaps + DigitCaps + FC | Unmentioned | Computational Intelligence Group, Basque University, "Hyperspectral remote sensing scenes" & VEDAS, ISRO, "AVIRIS-NG scenes" |
| | [28] | CNN1: Conv + BN + Relu + Dense layer<br>CNN2: Conv + Relu + BN + Dense layer<br>CAP1: Conv + Relu + BN + Conv + PrimaryCaps + DigitCaps<br>CAP2: Conv + BN + Relu + Conv + PrimaryCaps + DigitCaps | Duplicate removal + Noise Removal + Word Embedding | PaviaU (PU) & SalinasA (SA) |

| | | | | |
|---|---|---|---|---|
| | [29] | Pretraining: Conv + Logical Regression Classifier<br>Classification: Conv (Shallow Layers) + PrimaryCaps + ClassCaps + Masked ClassCaps + FC | Principal Component Analysis (PCA) + Splitting | Indian Pines |
| Time sequence data | [30] | Conv + PrimaryCaps + TrafficCaps | Time steps x Sensors Matrix | SETA EU project |
| | [31] | 1D CNN + ConvCaps + FC Caps + LSTM | 2D Matrix | UTSC-2016 |
| | [32] | Conv + PrimaryCaps + EmotionCaps + Three-layer feed-forward neural network | Multiband Feature Matrix | DEAP |
| | [33] | CNN: Conv + FC<br>CapsNet: Conv + PrimaryCaps + DigitCaps | Waveform framing + Feature extraction + Matrix transformation | Events recorded by between September 2017 and January 2019 in Huangtupo Copper and Zinc Mine |
| | [34] | Conv + PrimaryCaps + ClassificationCaps | Pearson correlation + Fisher's r-to-z transformation + Functional connectivity matrix | Data of 148 participants |
| Medical image | [35] | Conv + PrimaryCaps + ClassCaps + SegCaps | Grayscale image + Cropping + Hybrid filter, combination of Gaussian, anisotropic and a bilateral filter | Heidelberg SD-OCT imaging system |
| | [36] | Segmentation: U-net<br>Classification: Conv + PrimaryCaps + DigitCaps + FC | Pre-screening+ Marking + Segmentation + Unifying formats + Rotation operation | Kaggle competition |
| | [37] | Conv + PrimaryCaps + LabelCaps + FC | Resize + Data augmentation | Cohen |
| | [38] | Conv + PrimaryCaps + CancerCaps | Reinhard's method + Rotation + C70 | 400 hematoxylin and eosin (H&E) stained breast histology microscopy images |
| | [39] | Conv + PrimaryCaps + ClassCaps + FC | Unmentioned | Retrieval of Brain Tumors by Adaptive Spatial Pooling and Fisher Vector Representation |
| | [40] | Conv + Vector section + Output vectors | Region extraction + Noise removal + Resize + Gabor wavelet transform | face 95 and CASIA-Iris-Thousand |
| | [41] | Conv + PrimaryCaps + SEBlock + DigitalCaps + FC | Gray threshold segmentation + Connectivity analysis + left and right lung separation + Data reduction + Data enhancement | LIDC-IDRI |
| Text classification | [42] | Text representation module + GRU + ATT + Conv + PrimaryCaps + ClassCaps | Text representation | Chinese news data collected by Sogo lab |
| | [43] | Conv + PrimaryCaps + ConvCaps + FCCaps | Word2vec+Adam optimization algorithm with 1e-3 learning rate | MR, SST-2, Subj, TREC question dataset, CR & AGs news corpus |
| | [44] | Attention mechanism + IndRNN + FC + Sigmoid<br>Conv + PrimaryCaps + DigitCaps | Duplicate removal + Noise Removal + Word Embedding | Websites of Tianshan, Renming, etc. |
| Facial image | [45] | Conv + PrimaryCaps + MeCaps | Apex frame detection in a sequence data<br>including pixel calculation and noise removal | SMIC, CASME II & SAMM |
| City scene image | [46] | CNN + Conv + PrimaryCaps + FinalCaps | Unmentioned | UC Merced Land-Use, AID & NWPU RESISC45 |
| Agricultural image | [47] | Conv + PrimaryCaps + DigitCaps | Histogram equalization + Superpixel algorithm | Rice images captured by UAV |
| Fake image | [48] | Part of pretrained VGG-19 (CNN) + PrimaryCaps + OutputCaps | Frame sepapration + Segmentation | ILSVRC & FaceForensics++ |
| Word | [49] | Conv + PrimaryCaps + ClassCaps | MFCC and filter bank | MultiMNIST |
| Traffic sign | [50] | Conv + PrimaryCaps + DigitCaps | Image equalization + MSER segmentation + Normalization | GTSRB |
| Pedestrian image | [51] | Conv + PrimaryCaps + IntermediateCaps + AdvancedCaps | Unmentioned | CUHK01, CUHK03 & Market-1501 |
| Electricity patrol | [52] | Conv + PrimaryCaps + ImageCaps + FeatureCaps + FC | Hough Transform + Segmentation (CV) | 3000 images from a rail way by Beijing railway administration |

### C. Capsule Networks Used in Medical Image Processing

In the area of recognition and classification of medical images, deep learning models are found to be more efficient and accurate than humen. A few studies have applied capsule networks for processing various medical images and achieved satisfying results, showing the practicability of capsule networks in medical aspect. Koresh et al. [35] innovate an algorithm based on capsule networks for identifying the three major boundaries of the corneal layer. The work includes pre-processing, classification and segmentation. Zhang and Zhao [36] use a deep learning model for classification of cervical lesion images. The proposed method utilized CNNs for image segmentation to obtain the lesion part, then introduce a neural network similar to CapsNet for image recognition and classification. Toraman et al. [37] construct a convolutional capsule network for detecting COVID-19 disease through chest X-ray images. Iesmantas and Alzbutas [38] propose a deep learning strategy based on capsule networks for classifying four types of breast tissue biopsy images when hematoxylin and eosin staining is applied. Afshar et al. [39] have achieved 4 objectives in his work: design a modified capsule network for brain tumor classification; explore the overfitting problem based on a real MRI image dataset; compare the network's performance on whole brain images and segmented tumor images; develop a visualization paradigm for the output of the CapsNet to better explain the learned features. Jacob [40] puts forward a frame work that uses capsule networks for retina based biometric recognition. Feng et al. [41] apply capsule networks for low-dose computed tomography (CT) image recognition. These applications have shown capsule networks can assist medical image processing in a developable and stable way.

### D. Capsule Networks Used in Text Classification

Capsule networks are also widely used in text classification. The importance of different words in a sentence can be reflected by well-designed capsule networks, which have been proved by a few studies. Li et al. [42] propose a GRU-ATT-Capsule hybrid model, which utilizes gated recurrent unit (GRU) to extract contextual features and combines the attention mechanism (ATT) for studying the importance of the words in the text, then merges the these extracted feature, and uses capsule network to overcome the CNN's missing spatial information. Zhao et al. [43] explore capsule networks for text classification. Three strategies, Orphan Category, Leaky-SoftMax and Coefficients Amendment, are proposed to stabilize the dynamic routing process to mitigate the disturbance of some noise capsules. Wang et al. [44] propose combined a parallel model (Att_IndRNN_CapsNet) for filling missing arguments in Uyghur events based on attention-mechanical independent RNN and CapsNet. The attention-mechanical independent RNN is used for obtaining high-order features and CapsNet can provide contextual semantic features. Logical relation between texts can be accurately capture by capsule networks, making capsule networks an emerging solution for text processing.

### E. Capsule Networks Used in Other Senarios

Plentiful aspects are also concluded in the area of applying capsule networks, e.g., facial recognition, city scene recognition, agriculture image recognition, fake video recognition, text recognition, transportation sign recognition and pedestrian recognition. Quang et al. [45] innovate a simple but effective capsule network for micro-expression recognition. Zhang et al. [46] select a deep CNN model which was fully pretrained on ImageNet as a feature extractor, then fed the initial features into a newly designed CapsNet for remote sensing image scene classification. Li et al. [47] use rice images collected by UAV as the input data, and construct a capsule network to recognize rice images. Nguyen et al. [48] introduce a capsule network for detecting various kinds of attacks, from presentation attacks using printed images and replayed videos to attacks using fake videos. Xiong [49] utilizes capsule networks for solving overlapping digit recognition problem. Ren et al. [50] propose a Traffic Sign Recognition Algorithm (TSRA) based on capsule networks to recognize traffic signs. Simulation results demonstrate that TSRA is faster and more accurate than current methods. Cheng et al. [51] introduce a pedestrian recognition model based on the improved capsule networks. The model added 2 more convolutional layers and extended the capsule dimension based on the original CapsNet. Bian et al. [52] propose the CapsNet-CV algorithm for insulator damage identification. Capsule networks' application in electricity patrol is an outstanding example of machine learning assisting industrial projects. Applicability and universality of capsule networks are the footstone of capsule networks' future developments.

### F. Capsule Networks for Complex Tasks

Traditional CNN-based methods often tend to occupy massive memory and intricate structures if the data or task is complex. Yet capsule network can solve this problem by its stronger internal representation storage and routing information. Xi's paper [7] is one of the earliest literatures that apply capsule network on complex data classification. Experiments are carried out by testing different parameters and modification on the capsule network. Some of the new setting succeeded yet some failed. Xiong et al. [53] introduce the convolutional capsule layer to deepen CapsNet and achieve the best result of its time on CIFAR-10 dataset. Lalonde et al. [54] introduce a new capsule structure, deformable capsules, and SE-Routing, a novel dynamic routing algorithm. This modified structure is proved to be suitable for large-scale computer vision tasks and is the very first capsule network for object detection. Lalonde and Bagci [18] are the first to apply the capsule network on object segmentation. They propose to use a capsule network with locally-connected routing and deconvolutional capsules which substantially decreases parameter space. Zhang et al. [55] propose a CapsNet-based, GAN-combined structure for salient object detection; the structure is examined on a challenging dataset. A novel capsule conditional generative adversarial network (Caps-cGAN) with small number of parameters is proposed by Wang et al. [56] to construct the semi-supervised learning system and a novel joint semi-supervised loss function composed of unsupervised loss and supervised loss is proposed to train the model. Their work magnificently improved the situation when handling speckle noise in optical coherence tomography (OCT) images. Capsule networks' performance for complex tasks has well testified its potential and capacity compared with current other deep learning methods.

### G. Capsule Networks for Few-Shot Learning

Few-shot learning has always been a challenging aspect in machine learning, but is now a hotspot due to the rise of deep learning. Traditional deep learning methods often require massive data for parameter configuration and easily cause overfitting. Yet a few literatures have succeeded in this aspect with the help of capsule networks. Wu et al. [57] devise a prototypical network for few-shot learning with a new embedding structure applying capsule networks, a new triplet loss and an effective non-parametric classifier termed attentive prototypes. Neill [58] proposes the Siamese Capsule Networks which can process pairwise learning tasks. This new model shows strong baselines on both pairwise learning datasets. Geng et al. [59] reference the dynamic routing of capsule networks and use it in induction networks for few-shot texture classification, which outperforms the existing approaches. Anand et al. [60] utilize an auto-encoder to learn generalized feature embeddings from class-specific embeddings obtained from capsule network. Experimental results demonstrate the superiority and generalization ability of the proposed few shot learning pipelines. Capsule networks is without doubt a promising approach towards few-shot learning problems.

## IV. Discussion

When Hinton and his group propose the capsule networks [1, 2], they are aiming to eliminate the shortcomings of CNNs, such as the loss of equivariance and the loss of fine features. Capsule networks have outperformed some state-of-the-art convolutional neural networks on simple challenges like MNIST, MultiMNIST [2] or smallNORB [3], and shown good performance on some complex tasks as mentioned in Section III (F). Capsule networks have also been shown to reduce training time and reduce trainable parameters. However, some limitations of capsule networks have been already found in recent years. Peer et al. [61] prove that state-of-the-art routing procedures decrease the expressivity of capsule networks. Gu et al. [62] claim that capsule network is not more robust than convolutional network. They look into the special designs in capsule networks that differ from that of CNNs used for image classification. Their experiments reveal that some designs, which are thought critical to capsule networks, actually can harm its robustness, i.e., the dynamic routing layer and the transformation process. Moreover, deep capsule networks are slow in inference due to the high computational complexity and numerical instability of iterative routing mechanisms [63]. Therefore, in many aspects, CNNs still hold the advantages over capsule networks.

In the past 2 years, increasing attention has been drawn to transformers [64] in the deep learning research community. Transformer-based methods are reaching new state-of-the-art performance in more and more vision tasks [65]. It seems the self-attention based transformer module “is all we need”, and transformer presents the trend to replace the convolutional neural networks and dominate the deep learning technology. Some researchers are studying the connections and possibility of combining the capsule networks with transformers to take the full advantage of them.

Abnar [66] try to draw a connection between different components of transformers and capsule networks, and find a couple of functional similarities between them: both capsule networks and transformer architectures have a mechanism which allows the models to process the representations from a lower layer from different perspectives to compute the representation in the higher layer; in very loose terms, the pose matrix in capsule nets plays the role of key and query vectors in transformers. Duan et al. [67] propose the capsule-Transformer, which extends the linear transformation into a more general capsule routing algorithm by taking self-attention network as a special case of capsule network. So that the resulted capsule-transformer is capable of obtaining a better attention distribution representation of the input sequence. Mobiny et al. [68] propose a novel non-iterative routing strategy named self-attention routing that computes the agreement between the capsules in one forward pass. The self-attention routing utilizs a learnable inducing mixture of Gaussians to reduce the cost of computing pairwise attention values from quadratic to linear time complexity. Their design speeds up the networks and achieves superior accuracy. Very recently, Hinton [69] presents a single idea about representation which allows advances in capsule networks, transformer, CNNS, contrastive learning, neural fields, etc., to be combined into an imaginary system called GLOM to represent part-whole hierarchies in a neural network. He supposes that if GLOM can be made to work, it should significantly improve the interpretability of the representations produced by transformer-like systems when applied to vision or language.

We, as well as some other researchers, believe the biggest problem of capsule networks is the lack of good design and implementation. It need more time to evolve to excellent structure with fine implementation. In the future research, some measures can be taken to further boost the potential of capsule networks: 1) discover the complement parts of CNN, capsule networks and transformers, and take the full advantage of them; 2) devise better design of capsule networks based on inspiration from cognitive neuroscience and mathematical proof.

## V. Conclusion

This paper presents a relatively thorough literature review of capsule networks, including the evolution of the design, and a variety of applications of capsule networks. Although, capsule networks have achieved good performance on many simple computer vision tasks, their performance on complex tasks still need more validation. Capsule networks holds some advantages over CNNs in number of parameters, requirement of training data and representation of complex entities, but still suffer from low efficient and limited application range. It is promising to improve the performance of capsule networks by finding the complement parts of capsule networks, and existing learning systems and methods. We also think inspiration from cognitive neuroscience and mathematical proof will promote the advance in capsule networks.

## REFERENCES


[1] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming Auto-Encoders," in *ICANN 2011*, pp. 44-51.

[2] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *NeurIPS 2017*, Long Beach, California, USA, 2017.

[3] G. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *ICLR 2018*, Vancouver, BC, Canada, 2018.
[4] A. R. Kosiorek, S. Sabour, Y. W. Teh, and G. E. Hinton, "Stacked capsule autoencoders," in *NeurIPS 2019*, Vancouver, BC, Canada, 2019.
[5] A. T. Weiwei Sun, Boyang Deng, Sara Sabour, Soroosh Yazdani, Geoffrey Hinton, Kwang Moo Yi, "Canonical Capsules: Self-Supervised Capsules in Canonical Pose," in NeurIPS 2021.
[6] S. Sabour, A. Tagliasacchi, S. Yazdani, G. Hinton, and D. J. Fleet, "Unsupervised part representation by flow capsules," in *International Conference on Machine Learning*, 2021: PMLR, pp. 9213-9223.
[7] E. Xi, S. Bing, and Y. Jin, "Capsule Network Performance on Complex Data," *arXiv,* 2017.
[8] Y. Zhao, T. Birdal, H. Deng, and F. Tombari, "3D Point Capsule Networks," in *CVPR 2019*, pp. 1009-1018.
[9] Z. Chen and D. Crandall, "Generalized capsule networks with trainable routing procedure," *arXiv preprint arXiv:1808.08692,* 2018.
[10] J. E. Lenssen, M. Fey, and P. Libuschewski, "Group equivariant capsule networks," *arXiv preprint arXiv:1806.05086,* 2018.
[11] D. Wang and Q. Liu, "An optimization view on dynamic routing between capsules," in *ICLR 2018*, Vancouver, BC, Canada, 2018.
[12] D. Peer, S. Stabinger, and A. Rodriguez-Sanchez, "Training deep capsule networks," *arXiv preprint arXiv:1812.09707,* 2018.
[13] T. Hahn, M. Pyeon, and G. Kim, "Self-routing capsule networks," in *NeurIPS 2019*, Vancouver, BC, Canada, 2019.
[14] K. Ahmed and L. Torresani, "STAR-CAPS: Capsule networks with straight-through attentive routing," in *NeurIPS 2019*, Vancouver, BC, Canada, 2019.
[15] B. Mandal, R. Sarkhel, S. Ghosh, N. Das, and M. Nasipuri, "Two-phase Dynamic Routing for Micro and Macro-level Equivariance in Multi-Column Capsule Networks," *Pattern Recognition,* vol. 109, 2021.
[16] F. De Sousa Ribeiro, G. Leontidis, and S. Kollias, "Capsule routing via variational bayes," in *AAAI 2020*, New York, NY, United states, 2020.
[17] A. Mobiny, P. Yuan, P. A. Cicalese, and H. Van Nguyen, "Decaps: Detail-oriented capsule networks," in *23rd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2020*, Lima, Peru, 2020.
[18] R. LaLonde and U. Bagci, "Capsules for object segmentation," *arXiv preprint arXiv:1804.04241,* 2018.
[19] S. S. R. Phaye, A. Sikka, A. Dhall, and D. Bathula, "Dense and diverse capsule networks: Making the capsules learn better," *arXiv preprint arXiv:1805.04001,* 2018.
[20] J. Rajasegaran, V. Jayasundara, S. Jayasekara, H. Jayasekara, S. Seneviratne, and R. Rodrigo, "Deepcaps: Going deeper with capsule networks," in *CVPR 2019*, Long Beach, CA, United states, 2019.
[21] A. Hoogi, B. Wilcox, Y. Gupta, and D. L. Rubin, "Self-attention capsule networks for object classification," *arXiv preprint arXiv:1904.12483,* 2019.
[22] S. B. S. Bhamidi and M. El-Sharkawy, "Residual Capsule Network," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2019, pp. 0557-0560.
[23] N. Srivastava, H. Goh, and R. Salakhutdinov, "Geometric capsule autoencoders for 3d point clouds," *arXiv preprint arXiv:1912.03310,* 2019.
[24] Y. Zhao, T. Birdal, J. E. Lenssen, E. Menegatti, L. Guibas, and F. Tombari, "Quaternion Equivariant Capsule Networks for 3D Point Clouds," in *ECCV 2020*, Glasgow, United kingdom, 2020.
[25] H. Zhang *et al.*, "1D-Convolutional Capsule Network for Hyperspectral Image Classification," arXiv:1903.09834.
[26] X. Wang, K. Tan, Q. Du, Y. Chen, and P. Du, "Caps-TripleGAN: GAN-Assisted CapsNet for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 57, no. 9, pp. 7232-7245, 2019.
[27] P. V. Arun, K. M. Buddhiraju, and A. Porwal, "Capsulenet-Based Spatial–Spectral Classifier for Hyperspectral Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 12, no. 6, pp. 1849-1865, 2019.
[28] F. Deng, S. Pu, X. Chen, Y. Shi, T. Yuan, and S. Pu, "Hyperspectral Image Classification with Capsule Network Using Limited Training Samples," *Sensors (Basel),* vol. 18, no. 9, Sep 18 2018.
[29] J. Yin, S. Li, H. Zhu, and X. Luo, "Hyperspectral Image Classification Using CapsNet With Well-Initialized Shallow Layers," *IEEE Geoscience and Remote Sensing Letters,* vol. 16, no. 7, pp. 1095-1099, 2019.
[30] Y. Kim, P. Wang, Y. Zhu, and L. Mihaylova, "A Capsule Network for Traffic Speed Prediction in Complex Road Networks," *2018 Sensor Data Fusion: Trends, Solutions, Applications (SDF) IEEE,* 2018.
[31] H. Yao, P. Gao, J. Wang, P. Zhang, C. Jiang, and Z. Han, "Capsule Network Assisted IoT Traffic Classification Mechanism for Smart Cities," *IEEE Internet of Things Journal,* vol. 6, no. 5, pp. 7515-7525, 2019.
[32] H. Chao, L. Dong, Y. Liu, and B. Lu, "Emotion Recognition from Multiband EEG Signals Using CapsNet," *Sensors (Basel),* vol. 19, no. 9, May 13 2019.
[33] P. Peng, Z. He, L. Wang, and Y. Jiang, "Microseismic records classification using capsule network with limited training samples in underground mining," *Sci Rep,* vol. 10, no. 1, p. 13925, Aug 18 2020.
[34] T. Wang, A. Bezerianos, A. Cichocki, and J. Li, "Multikernel Capsule Network for Schizophrenia Identification," *IEEE Trans Cybern,* vol. PP, Dec 1 2020.
[35] H. J. D. Koresh, S. Chacko, and M. Periyanayagi, "A modified capsule network algorithm for oct corneal image segmentation," *Pattern Recognition Letters,* vol. 143, pp. 104-112, 2021.
[36] X. Zhang and S.-G. Zhao, "Cervical image classification based on image segmentation preprocessing and a

CapsNet network model," *International Journal of Imaging Systems and Technology,* vol. 29, no. 1, pp. 19-28, 2019.
[37] S. Toraman, T. B. Alakus, and I. Turkoglu, "Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks," *Chaos Solitons Fractals,* vol. 140, p. 110122, Nov 2020.
[38] T. Iesmantas and R. Alzbutas, "Convolutional Capsule Network for Classification of Breast Cancer Histology Images," in *Image Analysis and Recognition*, 2018, Chapter 97, pp. 853-860.
[39] P. Afshar, A. Mohammadi, and K. N. Plataniotis, "Brain Tumor Type Classification via Capsule Networks," *arXiv,* 2018.
[40] D. I. J. Jacob, "Capsule Network Based Biometric Recognition System," *Journal of Artificial Intelligence and Capsule Networks,* vol. 2019, no. 2, pp. 83-94, 2019.
[41] Y. Feng, W. Luyao, H. Wei, D. Guojun, and C. Jiarong, " Benign and Malignant Diagnosis of Pulmonary Nodules Based on SE-CapsNet," *Chinese Journal of Biomedical Engineering,* vol. 40, no. 1, p. 10, Feb. 2021.
[42] L. Ranran, L. Daming, L. Zheng, and C. Gaoxiang, "A Capsule Network Text Classification Method Integrating Stroke Features," *Computer Engineering,* 2021.
[43] W. Zhao, J. Ye, M. Yang, Z. Lei, S. Zhang, and Z. Zhao, "Investigating Capsule Networks with Dynamic Routing for Text Classification," *arXiv,* 2018.
[44] W. Xianxia, Y. Long, T. Shengwei, and W. Ruijin, " Missing Argument Filling of Uyghur Event Based on Independent Recurrent Neural Network and Capsule Network," *Acta Automatic Sinica,* vol. 47, no. 4, p. 10, 2021.
[45] N. V. Quang, J. Chun, and T. Tokuyama, "CapsuleNet for Micro-Expression Recognition," *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition,* 2019.
[46] W. Zhang, P. Tang, and L. Zhao, "Remote Sensing Image Scene Classification Using CNN-CapsNet," *Remote Sensing,* vol. 11, no. 5, 2019.
[47] Y. Li *et al.*, "The recognition of rice images by UAV based on capsule network," *Cluster Computing,* vol. 22, no. S4, pp. 9515-9524, 2018.
[48] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use os a Capsule Network to Detect Fake Images and Videos," *arXiv,* 2019.
[49] Y. Xiong, "Using Capsule Networks for Image and Speech Recognition Problems," *ProQuest LLC* 2018.
[50] R. Tiaojuan, C. Peng, C. Yourong, J. Jun, and Y. Jing, " Research on Traffic Sign Recognition Algorithm Based on Capsule Neural Network," *Automobile Technology*, 2020.
[51] C. Huan-xin, L. Wen-han, G. Zhan-guang, and Z. Zhi-hao, " Pedestrian Recognition in Complex Scenes Based on Capsule Network," *Computer Technoogy and Development,* vol. 31, no. 2, p. 75, 2021.
[52] B. Jianpeng, H. Jiaxing, Z. Shuai, H. Weijing, and G. Shichuang, " Insulator Damage Identification and Location Based on Improved Capsule Network," *Insulators and Surge Arresters,* no. 299, p. 7, 2021.
[53] Y. Xiong, G. Su, S. Ye, Y. Sun, and Y. Sun, "Deeper Capsule Network For Complex Data," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 14-19 July 2019 2019, pp. 1-8.
[54] R. LaLonde, N. Khosravan, and U. Bagci, "Deformable Capsules for Object Detection," *arXiv preprint,* Available: https://arxiv.org/abs/2104.05031.
[55] C. Zhang, F. Yang, G. Qiu, and Q. Zhang, "Salient Object Detection With Capsule-Based Conditional Generative Adversarial Network," in *ICIP 2019*, 2019, pp. 81-85.
[56] M. Wang *et al.*, "Semi-Supervised Capsule cGAN for Speckle Noise Reduction in Retinal OCT Images," *IEEE Transactions on Medical Imaging,* vol. 40, no. 4, pp. 1168-1183, 2021.
[57] F. Wu, J. S. Smith, W. Lu, C. Pang, and B. Zhang, "Attentive Prototype Few-Shot Learning with Capsule Network-Based Embedding," in *ECCV 2020*, 2020.
[58] J. O. Neill, "Siamese capsule networks," *arXiv preprint arXiv:1805.07242,* 2018.
[59] R. Geng, B. Li, Y. Li, X. Zhu, P. Jian, and J. Sun, "Induction networks for few-shot text classification," *arXiv preprint arXiv:1902.10482,* 2019.
[60] P. Anand, A. K. Singh, S. Srivastava, and B. Lall, "Few shot speaker recognition using deep neural networks," *arXiv preprint arXiv:1904.08775,* 2019.
[61] D. Peer, S. Stabinger, and A. Rodriguez-Sanchez, "Limitation of capsule networks," *Pattern Recognition Letters,* vol. 144, pp. 68-74, 2021.
[62] J. Gu, V. Tresp, and H. Hu, "Capsule Network is Not More Robust than Convolutional Network," *arXiv preprint,* Available: https://arxiv.org/abs/2103.15459.
[63] V. Mazzia, F. Salvetti, and M. Chiaberge, "Efficient-CapsNet: capsule network with self-attention routing," *Scientific Reports,* vol. 11, no. 1, p. 14634, 2021.
[64] A. Vaswani *et al.*, "Attention is all you need," in *NeurIPS 2017*, Long Beach, CA, United states, 2017.
[65] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030,* 2021.
[66] S. Abnar. "From Attention in Transformers to Dynamic Routing in Capsule Nets." https://samiraabnar.github.io/articles/2019-03/capsule (accessed May 8th, 2021).
[67] S. Duan, J. Cao, and H. Zhao, "Capsule-Transformer for Neural Machine Translation," *arXiv preprint arXiv:2004.14649,* 2020.
[68] A. Mobiny, P. A. Cicalese, and H. Van Nguyen, "Trans-Caps: Transformer Capsule Networks with Self-attention Routing," 2020.
[69] G. Hinton, "How to represent part-whole hierarchies in a neural network," *arXiv preprint arXiv:2102.12627,* 2021.