The 9th International Conference on Web Research
ICWR2023
May 3th-4th, 2023 ;Tehran , Iran

A Framework for Scalable Object Storage and Retrieval Considering Privacy Concerns: A Case Study on the Signature Detection

Mohammad Mehdi Daliri Khomami*[1], Ali Mohammad Saghiri[1], Alireza Rezvanian[2], Mohammad Reza Meybodi[1]
[1]Soft computing laboratory Department of Computer Engineering Amirkabir University of Technology, Tehran, Iran
M.daliri@aut.ac.ir, saghiri@aut.ac.ir, mmeybodi@aut.ac.ir
[2]Department of Computer Engineering, University of Science and Culture, Tehran, Iran
rezvanian@usc.ac.ir

***Abstract*— Nowadays, we are faced with a huge amount of time-changing private data generated in different ecosystems, including the Internet of Things, social networks, peer-to-peer networks, and e-commerce, to mention a few. The performance of traditional object storage and retrieval systems concerning response time is going to be degraded by a drastic increase in generated data. This problem becomes more challenging when a specific private object should be found. To address the mentioned problem, this paper suggests scalable object storage and retrieval. To do this, three well-known methods, content-addressable networks from peer-to-peer systems, vector space model from information retrieval systems, and siamese neural networks from neural networks theory, are involved in a framework to bring a scalable system to solve the mentioned problems. The suggested framework is unique to the best of our knowledge because there is no fusion among mentioned fields reported so far. A case study on signature detection is also conducted to evaluate the proposed framework. The results show that, compared to a centralized system, the proposed framework significantly decreases the response time for detecting a signature while maintaining the same accuracy. In addition, the proposed mechanism does not require the private objects to be sent to a central entity, which helps to alleviate privacy concerns.**

Keywords*— *Content-Addressable Network, Siamese Neural Networks, Vector Space Model, Scalability, Privacy.

I. INTRODUCTION

Recently, many ecosystems such as the Internet of Things (IOT), social networks, peer-to-peer, and e-commerce systems provide several facilities to end-users to generate different data on the Internet[1]. In this context, the end-users generate different content from different perspectives. Moreover, users' content generation causes the system to encounter large-scale data instead of rare data for processing. The rapid development of these contents has attracted significant attention from the research community from different aspects. One of the main challenges in this area is to provide efficient storage and retrieval methods for different objects to improve the system's response time. The performance of traditional object storage and retrieval systems from the response time aspect is going to be degraded by a drastic increase in generated data. This problem becomes more challenging for keeping data privacy, such as signature and fingerprint.

A number of studies [2]–[4] have been conducted to examine various properties of different storage and retrieval systems. Recent years have seen the widespread use of peer-to-peer systems, for example, as a means of file sharing, distributed computing, and information retrieval. Many techniques for indexing objects have been proposed in the litreature[5]. There is a single point of failure and performance bottlenecks with centralized indexing systems. By sending a query to every node, flood-based techniques like Gnutella consume lots of network bandwidth and CPU cycles. When using heuristic-based approaches for query probes, only a fraction of the nodes in the network is considered in order to reduce the number of probes to be used. As a general rule, these techniques may be categorized into three main categories: random walks, the use of summaries, and the grouping of nodes that have similar contents or share common interests.

Rhea and Kubiatowicz have presented a method for summarizing neighboring contents using Bloom filters, as per [2]. The technique operates based on a query that is sent only to appropriate objects in the vicinity with a high probability. In PlanetP [3], Bloom filters are utilized to summarize contents on each node and propagate them throughout the entire network. The idea of Routing Indices [4] offers a promising approach to finding appropriate objects, as opposed to Bloom filters. According to [5], nodes with similar content are grouped together to form a cluster. The algorithm commences with a random walk, but once it detects a group with matching content, it moves in a more deterministic manner. Cohen et al. [6] employ guide rules to classify nodes into an associative network based on data mining research. Sripanidkulchai et al. [7] develop a P2P network by connecting one node to another that benefits from previous queries. Replication is explored in [8] as one of the

policies to improve search efficiency. Consequently, "super-nodes" with high bandwidth are created to replicate the indices of several other nodes. As defined in [9], [10] suggests that maximizing the search size on successful queries can be achieved by minimizing the number of replicas per object. In hash-based systems [11], the hash function is converted into an ID, and the ID is utilized as the key in a DHT. These systems locate documents containing multiple query terms by intersecting the inverted lists [12]. This results in the algorithm's cost increasing proportionally with corpus size.

By considering the improvements mentioned in this section, one of the major drawbacks of this research is that signatures are naturally important data, and all proposed models are focused on centralized modeling, which can not preserve privacy and security. Among related studies, only [13] can produce a scalable search mechanism based on p2p that requires an intelligent system for indexing multimedia objects.

In this paper, scalable object storage and retrieval are suggested. In the proposed method, three well-known methods:

- Content-addressable networks from peer-to-peer systems
- Vector space model from information retrieval systems.
- Siamese neural networks from neural networks theory are combined in a framework to bring a scalable system to solve the mentioned problems

The suggested framework is novel because there is no fusion among mentioned fields reported so far. In order to examine the applicability of the proposed solution, a case study on signature detection is also suggested. The suggested solution does not need to the gathered private objects scattered in a network to a central entity, leading to some privacy concerns.

As an overview of the paper, section 2, provides an overview of the Peer-to-Peer system, Neural Networks, and Information Retrieval techniques. The proposed Siamese neural network for peer-to-peer is presented in section 3. The simulation results and a case study related to persian signature detection are investigated in section 4, and finally, future work and conclusions are provided in section 5.

II. PRELIMINARIES

The purpose of this section is to introduce the concepts that will be used in the proposed method. Following are the subsections that make up this part:

- Peer-to-Peer systems and content-addressable network
- Neural networks and Siamese Neural Networks
- Information retrieval technique and vector space model

For each parts, a short review are given to introduce the concept and then focus on the mentioned technique.

A. Peer-to-Peer systems and content-addressable network

A peer-to-peer system is a set of computers cooperating to share their resources, such as storage and computational power. These systems were introduced with Napster, and their file-sharing techniques have evolved during the last decade to share a vast amount of data for users. One of the well-known technology invented during the evolution of these networks was the content-addressable network, which is introduced in the next paragraph.

A content-addressable network (CAN) is a distributed hash table that is designed to scale, remain fault-tolerant, and self-organize to meet the requirements of various systems. To create a virtual multidimensional Cartesian coordinate space, an overlay network is formed using the CAN on a multi-torus. The resulting logical address is virtual, which implies that it is not dependent on the physical location or connectivity of nodes with each other. Points are detected in the space based on their coordinates. In this system, all nodes share the same dynamic coordinate space, and each node in the overall space possesses at least one unique zone. To join a CAN node, a node must first locate an overlay network node. Once the zone is split, nodes in the nearby area are identified, and their routing tables are updated.

B. Neural networks and Siamese Neural Networks

There has been a persistent effort to develop machines that can replicate human decision-making capabilities. This approach has led to the emergence of Neural Networks[14], Deep Learning[15], Learning Automata[16], and Cellular Learning Systems[17]. These tools have revolutionized algorithms in computer networks, social networks[18], and complex networks[19]. The use of deep learning networks has enabled significant advancements in performance due to their substantial computational power. However, these models have a significant drawback, unlike humans; they require a vast amount of ground truth samples containing correct answers, such as data, for the learning process. One way to address this issue is through one-shot learning, which formalizes the problem and enables the model to learn with a single reference point or sample and then identify the same instance in the testing data. Facial recognition systems commonly use one-shot learning, which is described through the use of Siamese Neural Networks (SNNs) in this sub-section. Siamese neural networks are constructed from two or more identical subnetworks that have the same parameters and weights in the same configuration. By comparing feature vectors, the network finds similarities between inputs. Traditional neural networks can predict multiple classes, but it becomes challenging to add or remove existing classes from the data. Therefore, we must retrain the entire dataset to update the neural network. SNNs, on the other hand, learn and train a similarity function to determine if two images are similar, allowing for classification of new data classes without retraining the network. Fig.1 illustrates the entire structure of the Siamese neural network. SNNs are commonly used in many domains, such as signature detection, for one-shot learning.

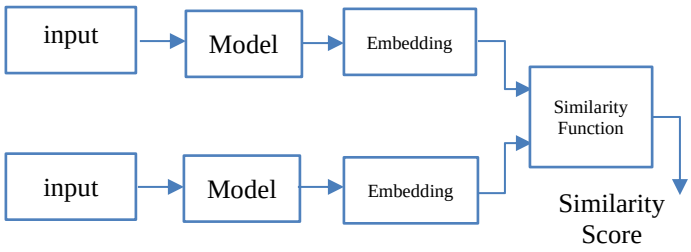

Fig. 1. The Architecture of Siamese nural network

A SNN is composed of two identical neural networks that share the same architecture and weights. This model contains multiple instances of the same architecture. To detect signatures, SNNs go through four stages of training. First, a collection of forged and similar signature images is gathered. The second stage separates the data into two parts: Signature S1i, which is a duplicate of S2i, but the remaining signatures in the second batch are not duplicates of S1i. The third stage involves inputting the batch file from the previous stage into a Convolutional Neural Network (CNN), which generates a 1-dimensional feature vector, and computing the cosine similarity between them. The above steps are repeated in the fourth stage for forward propagation and updating of neural network weights. A Triplet Loss can be used to train a Siamese network [20].

C. Information retrieval technique and vector space model

Hash tables are an efficient way to map keys to values, and they find widespread use in software systems. Content-Addressable Networks (CAN) are distributed hash tables used in many large-scale distributed systems. Peer-to-peer file-sharing services like Napster and Gnutella can benefit from the functionality of CAN. In these systems, files are not stored on a central server, but on end-user machines (peers) and are directly transferred between peers. Peer-to-peer systems have become increasingly popular, with Napster, for example, being downloaded by over 50 million users since its introduction in mid-1999, making it the fastest-growing application on the Web in December 2000. Other recent file-sharing systems include Scour, FreeNet, Ohaha, Jungle Monkey, and MojoNation.

Numerous peer-to-peer structures are available, but scalability is a major issue for most of them. Napster, for example, stores an index of all available files on a central server within its community. To retrieve a file, a user queries the central server with the name of the file and obtains the IP address of the machine that has the requested file. The user's machine then directly downloads the file from the server. Although Napster uses a peer-to-peer communication model, the process of locating a file remains centralized, which is expensive and vulnerable since there is a single point of failure. Gnutella addresses this issue by decentralizing the process of locating files.

Users on Gnutella networks form a mesh at an application level where file requests with a specific scope are broadcasted. However, this flooding approach on every request is not scalable and may not be able to find real content due to curtailing the flooding at some point. Hence, we asked ourselves: Can we create a scalable peer-to-peer file distribution system? It was soon realized that the indexing scheme, which maps file names from their location in the system to their names, is critical to any peer-to-peer system. In essence, peer-to-peer file transfers can be scaled, but locating the right peer is challenging. Thus, to have a scalable peer-to-peer system, the indexing mechanism must also be scalable. This paper proposes a specific design for such indexing systems, called Content-Addressable Networks (CANs).

III. PROPOSED FRAMEWORK: SNN-P2P

In the proposed framework, a large number of peers are organized into an overlay network to organize the information retrieval system in a self-organized manner. In this system, all peers execute a similar set of functions regarding storing and retrieval of objects. Each peer is able to publish object indices or submit queries for locating an object. The overlay network of the proposed framework can be constructed over any underlying network that provides peer-to-peer communication over the internet. Therefore, there is no limitation on the underlying network technology, and hence the underlying network can be sensor networks, Internet of Things, and edge to edge computing in the edge-cloud ecosystem. In the next paragraph, the rationale behind the proposed framework is explained.

In this network, the objects are stored and retrieved based on their features extracted by SNN. In a peer, to find an object, the object is converted to a Vector based on SNN executed in that peer. Then we can extract similar objects by executing a neighborhood query in the network. In other words, since peers are connected based on their object similarity, finding similar objects based on cosine similarity is converted to a simple neighborhood search in the network. This trick leads to the proposed system easily scaling up after increasing the new contents. In the next paragraph, a snapshot of the proposed system is explained for more clarification.

Fig.2 shows a snapshot of the system, Fig.3 depicts the process of indexing objects of $peer_i$ the network and Fig.4 illustrates the process of finding object x in every.

It should be noted that CAN protocol should be implemented in the network as a required infrastructure. The proposed framework can execute storing and retrieval processes fully distributed and self-organized. There is no external entity or manager to manage the network functionality.

A. Case study: Signature detection system

In order to utilize the proposed framework, signatures data were used. This is because of three reasons as explained as below:

- The first reason is that the information of signatures require many privacy concerns.
- The second reason is that the size of signatures data is increasing day by day.
- The third reason is that the proposed framework does not need to move the place of signatures from their

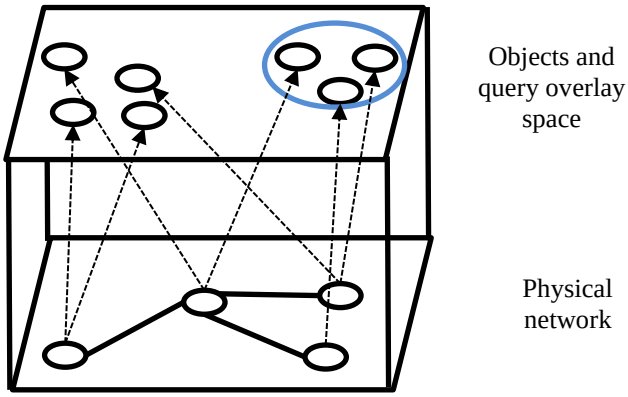

Fig. 2. Overview of SNN-P2P Searching system

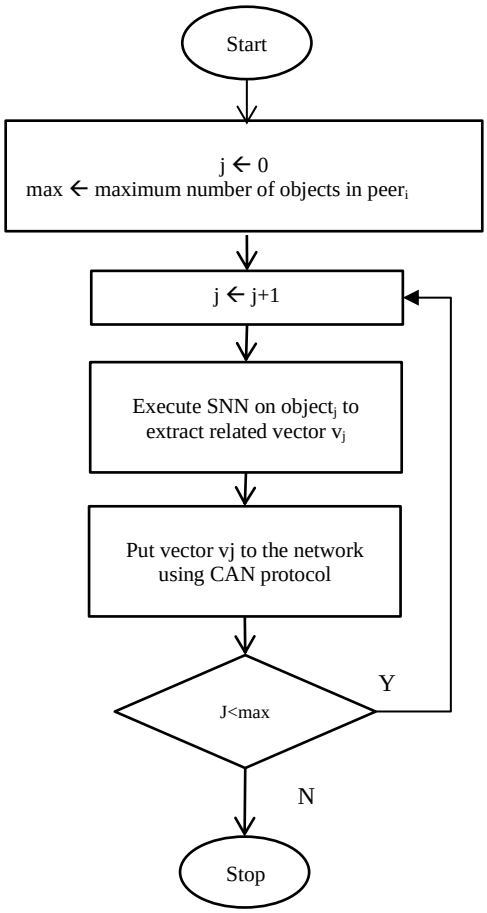

Fig. 3. flowchart for indexing objects in peer$_i$

own position to another place, and we can suggest a detection process in a fully distributed manner.

Fig.5 shows the case study on which this study investigate as an application.

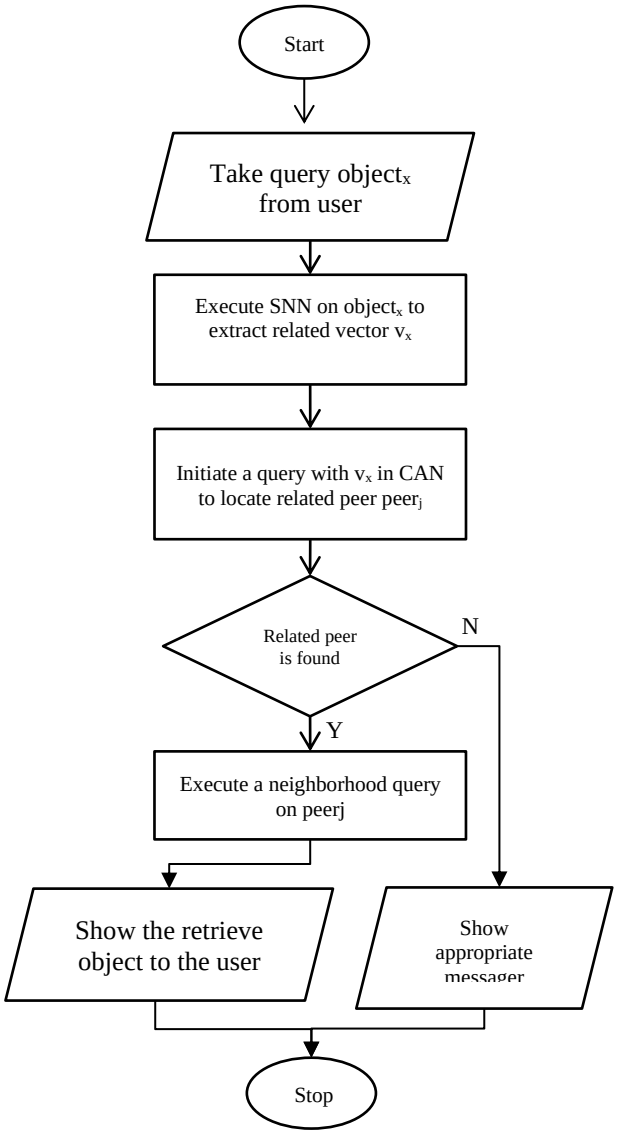

Fig. 4. flowchart for finding object$_x$

IV. Simulation Results

In order to evaluate the performance of the proposed algorithm, computational experiments are performed on the 100000 persion signature image which is distributed in P2P networks randomly among 10000 peers. Moreover, we compared the proposed algorithm when the architecture of the Siames network is central in comparison with the distributed proposed algorithm that utilizes siames network in fully distributed manner. Response time includes the time taken to transmit the inquiry, process it by the computer, and send the first response back to the user. In addition, the obtained result is reported in terms of the response time and also visited node to show the efficiency of the proposed algorithm. The result is presented in Fig.6 where the X-axis is the number of nodes and Y-axis is the percentage of response time.

Fig. 5. A case study Scheme for the SNN-P2P

From the result, we may conclude that as the size of the signature image in the network proceeds, the response of the algorithm drastically decreases. Moreover, to show the efficiency of the proposed algorithm, we have plotted the number of return images versus total visited nodes. The obtained result is depicted in Fig.7

The result shows that the number of visited nodes grows quickly, while the accuracy remains the same. More clearly, when the user requests 15 signatures, 5.9 nodes need to be searched to find relevant signatures. When the number of returned documents increases to 960, only 0.47 nodes need to be searched to find one relevant signature. We suggest using 15 as the default number of returned signatures is appropriate.

V. Conclustion

In this paper, a distributed system for object retrieval was proposed. The object used in the proposed framework can be an image, voice, or also text. In order to study the applicability of the proposed framework, a system for signature detection was developed. The proposed framework has several benefits as explained below:

- Because of the high privacy of information related to signatures, the proposed detection systems will have many applications in banking systems and governmental organizations.
- Because of the high computational power required for organizing a detection system for signatures, the proposed system will have many use cases because of avoiding gathering data in a center and imposing a huge amount of computation.

Regarding response time and visited nodes, the proposed algorithm performed better than a centralized one which may be used to support the mentioned claimes. As future work, the proposed framework can be used in edge-to-edge computation, which drastically decreases the required computational power related to establishing cloud and fog systems.

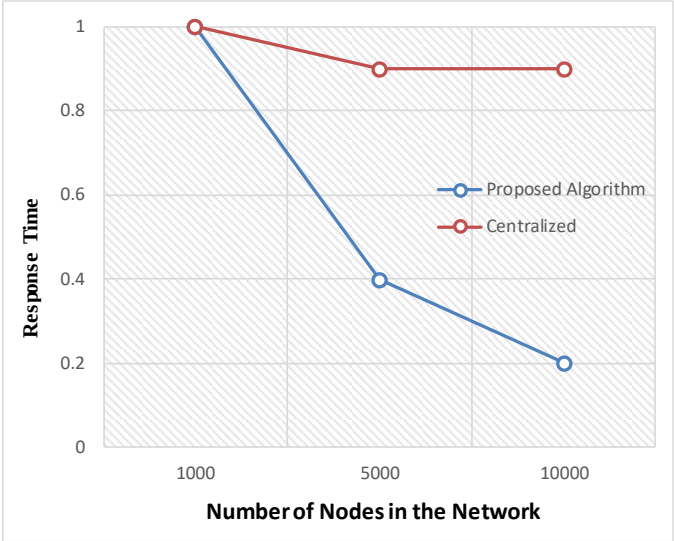

Fig. 6. Performance of the proposed algorithm in terms of response time.

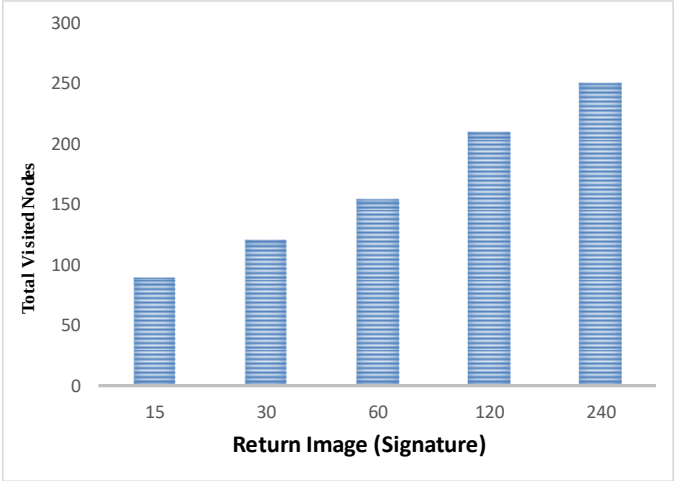

Fig. 7. The visited Nodes per returned signature

VI. References

[1] R. Ameri, M. R. Meybodi, and M. M. Daliri Khomami, "Cellular Goore Game and its application to quality-of-service control in wireless sensor networks," *The Journal of Supercomputing*, pp. 1–48, 2022.

[2] F. Concas, P. Xu, M. A. Hoque, J. Lu, and S. Tarkoma, "Multiple set matching with bloom matrix and bloom vector," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 14, no. 2, pp. 1–21, 2020.

[3] F. M. Cuenca-Acuna and T. D. Nguyen, "Text-based content search and retrieval in ad-hoc p2p communities," in *International Conference on Research in Networking*, Springer, 2002, pp. 220–234.

[4] A. Crespo and H. Garcia-Molina, "Routing indices for peer-to-peer systems," in *Proceedings 22nd international conference on distributed computing systems*, IEEE, 2002, pp. 23–32.

[5] M. F. Schwartz, *A scalable, non-hierarchical resource discovery mechanism based on probabilistic protocols*. Citeseer, 1990.

[6] E. Cohen, A. Fiat, and H. Kaplan, "Associative search in peer to peer networks: Harnessing latent semantics," *Computer Networks*, vol. 51, no. 8, pp. 1861–1881, 2007.

[7] K. Sripanidkulchai, B. Maggs, and H. Zhang, "Enabling efficient content location and retrieval in peer-to-peer systems by exploiting locality in interests," *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 1, pp. 80–80, 2002.

[8] J. Liang, R. Kumar, and K. W. Ross, "The FastTrack overlay: A measurement study," *Computer Networks*, vol. 50, no. 6, pp. 842–858, 2006.

[9] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker, "Search and replication in unstructured peer-to-peer networks," in *Proceedings of the 16th international conference on Supercomputing*, 2002, pp. 84–95.

[10] E. Cohen and S. Shenker, "Replication strategies in unstructured peer-to-peer networks," *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 4, pp. 177–190, 2002.

[11] J. Li, B. T. Loo, J. M. Hellerstein, M. F. Kaashoek, D. R. Karger, and R. Morris, "On the feasibility of peer-to-peer web indexing and search," in *International Workshop on Peer-to-Peer Systems*, Springer, 2003, pp. 207–215.

[12] M. W. Berry, Z. Drmac, and E. R. Jessup, "Matrices, vector spaces, and information retrieval," *SIAM review*, vol. 41, no. 2, pp. 335–362, 1999.

[13] C. Tang, Z. Xu, and S. Dwarkadas, "Peer-to-peer information retrieval using self-organizing semantic overlay networks," in *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, 2003, pp. 175–186.

[14] J. He, M. Xiao, J. Zhao, Z. Wang, Y. Yao, and J. Cao, "Tree-structured neural networks: Spatiotemporal dynamics and optimal control," *Neural Networks*, 2023.

[15] D. Huang, J. Liu, Y. Shi, C. Li, and W. Tang, "Deep polyp image enhancement using region of interest with paired supervision," *Computers in Biology and Medicine*, p. 106961, 2023.

[16] M. M. D. Khomami, M. A. Haeri, M. R. Meybodi, and A. M. Saghiri, "An algorithm for weighted positive influence dominating set based on learning automata," in *Knowledge-Based Engineering and Innovation (KBEI), 2017 IEEE 4th International Conference on*, IEEE, 2017, pp. 0734–0740.

[17] M. M. D. Khomami, A. Rezvanian, and M. R. Meybodi, "A new cellular learning automata-based algorithm for community detection in complex social networks," *Journal of computational science*, vol. 24, pp. 413–426, 2018.

[18] M. M. D. Khomami, A. Rezvanian, N. Bagherpour, and M. R. Meybodi, "Irregular cellular automata based diffusion model for influence maximization," in *2017 5th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, IEEE, 2017, pp. 69–74.

[19] M. M. D. Khomami, A. Rezvanian, and M. R. Meybodi, "Distributed learning automata-based algorithm for community detection in complex networks," *International Journal of Modern Physics B*, vol. 30, no. 8, p. 1650042, 2016.

[20] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 459–474.