MAMA Net: Multi-Scale Attention Memory Autoencoder Network for Anomaly Detection

Yurong Chen, *Member, IEEE*, Hui Zhang, *Member, IEEE*, Yaonan Wang, Yimin Yang, *Senior Member, IEEE*, Xianen Zhou, and Q. M. Jonathan Wu, *Senior Member, IEEE*

***Abstract*— Anomaly detection refers to the identification of cases that do not conform to the expected pattern, which takes a key role in diverse research areas and application domains. Most of existing methods can be summarized as anomaly object detection-based and reconstruction error-based techniques. However, due to the bottleneck of defining encompasses of real-world high-diversity outliers and inaccessible inference process, individually, most of them have not derived groundbreaking progress. To deal with those imperfectness, and motivated by memory-based decision-making and visual attention mechanism as a filter to select environmental information in human vision perceptual system, in this paper, we propose a Multi-scale Attention Memory with hash addressing Autoencoder network (MAMA Net) for anomaly detection. First, to overcome a battery of problems result from the restricted stationary receptive field of convolution operator, we coin the multi-scale global spatial attention block which can be straightforwardly plugged into any networks as sampling, upsampling and downsampling function. On account of its efficient features representation ability, networks can achieve competitive results with only several level blocks. Second, it's observed that traditional autoencoder can only learn an ambiguous model that also reconstructs anomalies "well" due to lack of constraints in training and inference process. To mitigate this challenge, we design a hash addressing memory module that proves abnormalities to produce higher reconstruction error for classification. In addition, we couple the mean square error (MSE) with Wasserstein loss to improve the encoding data distribution. Experiments on various datasets, including two different COVID-19 datasets and one brain MRI (RIDER) dataset prove the robustness and excellent generalization of the proposed MAMA Net.**

***Index Terms*— Anomaly detection, COVID-19 diagnose, attention mechanism, hash coding, memory autoencoder.**

I. Introduction

ANOMALY detection indicates the problem that finding patterns that are non-conforming to expected behavior. It has been extensively researched in the computer vision field because of its potential applications in medical image diagnoses [1], video surveillance [2], and network (social network, finance) analysis [3], *etc*. Recently, with the rapidly increasing demand for effective and efficient public health anomaly detection and real-time surveillance, developing automatic abnormal events detection system is one critical task which can tremendously alleviate labor-intensive work and non-stop human attention. For instance, the global pandemic Coronavirus Disease 2019 (COVID-19) has spread rapidly across the world [4]. For supplement the low sensitivity of the reverse-transcription polymerase chain reaction (RT-PCR) [5], automatically efficient computer-aided anomaly detection using X-rays or computed tomography (CT) offers great potential for tackling COVID-19.

Since the first statistics community study for anomalies detection was finished as early as the 19^{th} century [6], over time, a spectrum of anomaly detection methods have been developed, including the one-class classification algorithm [7], [8], such as support vector machines (SVMs) [7], and neural networks one-class classification methods: deep one-class (DOC) [8] and so on. However, the prerequisite for all these successes is the availability of corpora normal labels, which is difficult to define the encompasses of the normal item. Meanwhile, specifying all novel examples is impossible so that their further variants are inherently limited by a low recall [9]. In addition, compared with normal instances, abnormalities are rare, resulting in unbalanced positive and negative sample rates when training those algorithms; Also labeling is time-consuming and labor-consuming, especially for those high-demanding works like medical images diagnoses, which deters their deployment in many real-word applications.

On the other hand, unsupervised anomaly detection approaches [9], [10] aim to learn a pattern recognition that only given normal data then detect abnormal examples that do not conform to the normal profile. Although some studies can achieve satisfying performance in some particular scenarios,

Manuscript received October 25, 2020; revised November 22, 2020; accepted December 13, 2020. Date of publication December 16, 2020; date of current version March 2, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61971071, Grant 62027810, and Grant 61701047; in part by the National Key Research and Development Program of China under Grant 2018YFB1308200; in part by the Hunan Key Laboratory of Intelligent Robot Technology in Electronic Manufacturing under Grant IRT2018009; in part by the Hunan Key Project of Research and Development Plan under Grant 2018GK2022; and in part by the Changsha Science and Technology Project under Grant kq1907087. *(Corresponding author: Hui Zhang.)*

Yurong Chen, Hui Zhang, Yaonan Wang, and Xianen Zhou are with the National Engineering Laboratory of Robot Visual Perception and Control Technology, School of Robotics, Hunan University, Changsha 410082, China (e-mail: chenyurong1998@outlook.com; zhanghuihby@126.com, yaonan@hnu.edu.cn; zhouxianen@hnu.edu.cn).

Yimin Yang is with the College of Computer Science, Lakehead University, Thunder Bay, ON P7B 5E1, Canada (e-mail: yyang48@lakeheadu.ca).

Q. M. Jonathan Wu is with the College of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (e-mail: jwu@uwindsor.ca).

Digital Object Identifier 10.1109/TMI.2020.3045295

1558-254X © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.

the inaccessible training and inference process due to lack of constraints, human supervision with labels and generalization problem lead to a significant obstacle for carrying out a versatile framework [11]. The inconsistent results in high-dimensional data spaces, due to the "Curse of dimension" [12], which is another vital challenge of those approaches. Moreover, in [12], a semi-supervised autoencoder is treated as a feature extractor which outputs representations for traditional classifier such as one-class SVM. In general, autoencoder is trained to minimize the reconstruction error between normal samples input and output of decoder and enlarge the error between anomaly and decoder output. However, due to the "excellent" generalization ability that the network can only learn an ambiguous model, the reconstruction of abnormalities is also undistinguished with corresponding inputs by deep autoencoder. Moreover, the presumption that abnormalities produce high reconstruction error is debatable because abnormal samples are inaccessible during the training procedure and the reconstruction processes for anomaly cases are unpredictable [13], [14].

Albeit fruitful progress has been made in the last several years for working out those challenges, such as [9] combines variational Bayes and neural networks to obtain a commendable generation model. Vercruyssen *et al.* [13] propose a constrained-clustering-based approach for anomaly detection and [14] proposes a deep autoencoder with a memory module that the autoencoder reconstructs the most relevant instance of the input in the memory module. Despite of those techniques introduce supervision to guide their network, they are hampered by the limited receptive field of single convolution operator and vanishing gradient problem from the deep network due to the competing depth and width [15], which results poor perceptual quality and cannot fully exploit the potential of global information. In detail, the shallow layers cannot be trained usable to encode and reconstruct the input image well [16].

Inspired by interactions between visual processing and visual attention as a filter to select environmental information for learning, as well as the contribution of visual attention to memory [17], in this paper, we coin the multi-scale attention block that can be used in encoder and decoder for feature extracting and data reconstruction which can replace the single convolutional layer and transpose convolutional layer in traditional deep autoencoder network. The multi-scale attention block can achieve sampling, upsampling, and downsampling straightforwardly with combining channel attention layer which can mitigate the channel information loss and focus on the discriminative feature maps and pixel patch attention layer which has the ability to identify specific locations information within their footprint [18]. In the end, the model can outperform or match state-of-the-art networks with two or three blocks, which is beneficial to convergence and inference speed with 1.3× and 2.7× fewer FLOPs than ResNet101 with 7.6 GFLOPs and VGG16 with 15.5 GFLOPs backbone network (our backbone with 5.7 GFLOPs).

In addition, to mitigate the drawback of the traditional autoencoder network and those variant algorithms [9], [14] that due to lack of constraints and human intervention, it's observed that they can only learn an ambiguous model and it also reconstructs anomalies "well", in other words, it cannot produce higher reconstruction error than normal instances, leading to the miss detection of anomalies. To mitigate this challenge, motivated by [14] that human choices are shaped by awareness of past experiences and anticipation of future possibilities, furthermore, to bridge the gap of unstable similarity measurement and slow image retrieval methods, [19] shown superior improvements of similarity search and the work [20], [21] that proves semantic image hashing as a potent tool for image retrieval, in this paper, we propose a generalized deep autoencoder with a hash memory module. The difference between traditional autoencoder and MAMA Net and its main process are shown in Fig. 1. In the training, the parameter of hash memory module with a fixed number of memory slots is updated. After only training on the normal dataset, the learned hash table memory module is settled. Therefore, the normal testing data can retrieve similar hashing code and then get a low reconstruction error. On the other hand, the anomaly has to retrieval the nearest normal neighbor in hash memory module which will make a high reconstruction error.

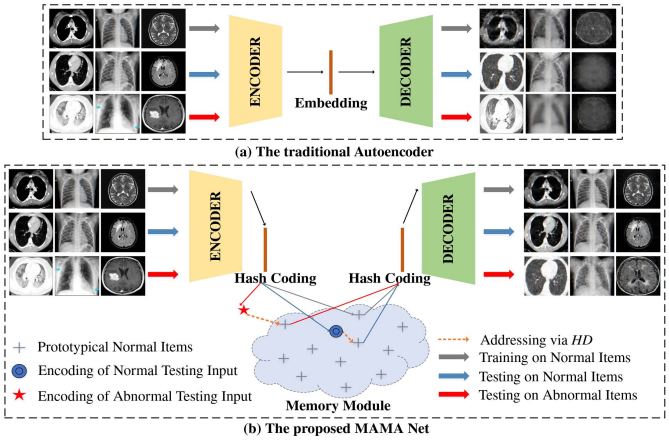

Fig. 1. The difference between the traditional autoencoder network (top) and our proposed MAMA Net (bottom) and its main components. Both are trained on normal items as the gray line and tested on new normal cases as the blue line and anomaly as the red line. In the training phase, the memory module is updated with prototypical normal items as plus sign. And in the testing process, the memory module is fixed and both normal cases and anomaly are addressed via hamming distance.

In conclusion, we proposed a generalized attention hash addressing memory autoencoder based framework named as MAMA Net for automatic anomaly detection, especially for COVID-19 detection. The main contributions of this paper are summarized as follows:

1): A hash addressing memory module is designed for fast retrieving the most relevant item and more appropriate produce high reconstruction error of the anomaly.

2): A multi-scale attention block with combining pixel patch attention layer and channel attention layer is proposed for replacing the convolution layer and transpose convolution layer, which can achieve sampling, downsampling, and upsampling function.

3): Couples Wasserstein distance (Earth-Mover distance) for the sake of same data distribution with mean square error (MSE) to build a new loss function.

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.

We apply the proposed MAMA Net on various comprehensive datasets from different applications. Experiments have proven its excellent robustness, high effectiveness, and generalization ability.

II. Related Work

A. Feature Extraction

To reduce the redundant information of input, feature extraction is essential in pattern recognition and image processing. The inappropriate feature may cause an algorithm to overfit and generalize poorly to new samples [22]. The traditional algorithm [23] is not enough robust in complex or crowded scenes, especially dealing with high dimensional data. In the recent years, feature extraction methods attracted a lot of attention, such as with the tremendous growth of the neural network, deep autoencoder [25] was proposed to learn efficient data representations in an unsupervised manner. Moreover, in order to accommodate more flexible modeling and more complex datasets representation, attention mechanisms [24] became the basic building block of most state-of-the-art architectures. Convolutional neural networks (CNNs) shows impressive success in image processing [25] due to its feature extraction ability. Instead of using artificial neural networks in deep autoencoder, [26] presents a convolutional autoencoder (CAE) for unsupervised feature learning.

B. Anomaly Detection

In the introduction section, we briefly go through the menagerie of traditional anomaly detection work and their dilemmas [7], [8], [12], [13]. In addition, Luo *et al.* [2] adopts stacked RNN with iteratively update the sparse coefficients to detect anomalies. However, this algorithm demands a robust and stable recurrent mechanism to encourage the model to generate large reconstruction error on the anomalies. Moreover, [8] purposes two loss functions, compactness loss and descriptiveness loss, to facilitate deep one-class (DOC). With recent achievements in the applications of deep neural networks (DNNs), memory augmented DNNs have attracted researchers' interest in increasing the robustness of networks [14], [34]. Gong *et al.* [14] propose a memory-augment autoencoder to mitigate the drawback that the reconstruction error of anomalies is inconspicuous. [34] presents a generative model with an attention mechanism to capture the local information. Lu *et al.* [35] use a layerwise procedure to train autoencoder models that can capture the intrinsic difference between outliers and normal instances.

C. Artificial Intelligence for COVID-19

The global outbreak of COVID-19 substantially increases doctors and and other medical workers' workload, which make automated anomaly detection for COVID-19 urgently needed. For supplement the low sensitivity of the RT-PCR, automated efficient computer-aided detection using X-Ray and computed tomography (CT) offers great potential for tackling COVID-19. Radiological imaging testing is of considerable practical importance especially in the early stages of COVID-19 [5]. Typical deep learning framework: Classification and segmentation method depends on DenseNet [27]; Unet is adopted in [28] as a weakly-supervised model to predict the COVID-19 infectious probability and in [29] to develop detection Coronavirus. Moreover, [30] modifies the Inception transfer-learning model for classifying COVID-19 patients. Xi *et al.* [31] develop a dual-sampling attention network to automatically diagnose COVID-19. And [32] performs COVID-19 detection in a weakly-supervised manner. Inf-Net [33] is proposed to automatically identify infected regions from chest CT slices. Although significant progress achieved, those methods only focus on certain modality and a generalized and comprehensive framework for COVID-19 detection is imperative.

III. Multi-Scale Attention Hash Memory Autoencoder

A. Overview

The proposed MAMA Net mainly involves three parts (see Fig. 2): an encoder network for learning the latent representations of input data; a hash memory module that given a query tensor from the encoder, it can retrieve the most homogeneous value tensor via the hamming distance of hash coding; a decoder network for reconstructing the value tensor from the memory module. During the training process, the encoder, decoder network, and memory module are trained to optimize the reconstruction error. Different from previous works [13], [14] that only using MSE as loss function, the Earth-Mover distance is joint for similar data distribution. The hash memory module is simultaneously optimized and updated during the training procedure. In the testing phase, the memory module is fixed. So that given a normal sample, it can retrieve a similar normal item in the memory module and result in a small reconstruction error. On the contrary, an anomaly instance incurs high error due to it has to retrieve the most relevant normal item. Instead of the commonly stacking convolutional layer, we propose the multi-scale attention block which can greatly reduce the depth of neural networks and achieve better feature extraction. In Section III-B, the structure of the multi-scale attention block is provided. In Section III-C, the hash memory module will be discussed. And the whole framework can be seen in Section III-D. Table I defines the symbols used in this article.

B. Multi-Scale Attention Block

To mitigate the limited receptive field of invariably local operators of single convolutional layer and deconvolutional layer, the multi-scale attention block can fuse global information effectively and efficiently. Unlike the self-attention transformer [24] only generates outputs with same feature sizes as the input, the proposed multi-scale attention block can be employed for regular sampling, downsampling and upsampling that can outputs any arbitrary dimensions (see Fig. 3) with a novel combination of pixel patch attention layer and channel attention layer. Specifically, we define Same Attention (SA) block for regular convolution with generates same size feature maps; Down Attention (DA) block for downsampling with outputs half size and Up Attention (UA) block for upsampling with produce double dimension. Even those three types of block generate different dimension outputs, they share similar

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.

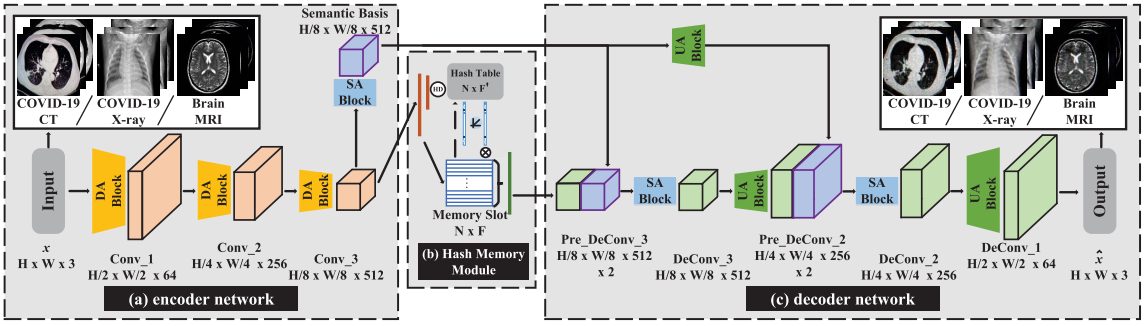

Fig. 2. The structure of proposed MAMA Net: (a) an encoder network for encoding input data; (b) a hash memory module that given a query tensor from the encoder, it can retrieve the most homogeneous value tensor via the hamming distance of hash coding; (c) a decoder network for reconstruction.

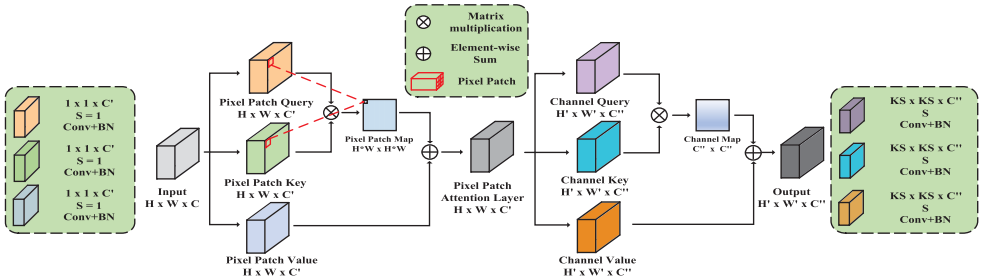

Fig. 3. The framework Multi-Scale Attention Block consists of two parts: (i) pixel patch attention layer (ii) channel attention layer. All SA, DA, and UA block can be realized with this framework only need to change the kernel size *KS* and stride *S*.

TABLE I
SYMBOLS DEFINITION

Symbol	Definition
SA	The Same Attention block.
DA	The Down Attention block.
UA	The Up Attention block.
$\mathcal{P}_\mathcal{Q}$	The pixel patch query tensor.
$\mathcal{P}_\mathcal{K}$	The pixel patch key tensor.
$\mathcal{P}_\mathcal{V}$	The pixel patch value tensor.
$Conv1_{C'}(\cdot)$	The convolution operator with kernel size = 1×1, stride = 1 and output C' channel feature maps.
$Patch_Q$	The particular query patch tensor.
$Patch_K^T$	The transpose of particular key patch tensor.
$\mathcal{C}_\mathcal{Q}$	The channel query tensor.
$\mathcal{C}_\mathcal{K}$	The channel key tensor.
$\mathcal{C}_\mathcal{V}$	The channel value tensor.
$G(\cdot)$	The generator of channel query tensor.
M	The memory slot matrix.
N	The number of memory slot.
F	The dimension of each memory item.
HS	The hash addressing matrix.
$r(\cdot)$	The hash mapping function.
z	The encoding latent representation of input.
w_{cos}	The similarity coefficients of cosine.
HD	The hamming distance.
w_{HD}	The similarity coefficients of hamming distance.
MES	The mean square error loss.
EM	The Wasserstein (Earth-Mover) distance loss.

design ideas and structures as Fig. 3 shown. Given inputs of multi-scale attention blocks are denoted as $Input \in \mathbb{R}^{H\times W\times C}$, the first process is to get pixel patch query tensor $\mathcal{P}_\mathcal{Q}$, pixel patch key tensor $\mathcal{P}_\mathcal{K}$ and pixel patch value tensor $\mathcal{P}_\mathcal{V}$. Here, we adopt three three convolution layers as

$$\begin{aligned}\mathcal{P}_\mathcal{Q} &= Conv1_{C'}(Input) \in \mathbb{R}^{H\times W\times C'},\\ \mathcal{P}_\mathcal{K} &= Conv1_{C'}(Input) \in \mathbb{R}^{H\times W\times C'},\\ \mathcal{P}_\mathcal{V} &= Conv1_{C'}(Input) \in \mathbb{R}^{H\times W\times C'},\end{aligned} \tag{1}$$

where $Conv1_{C'}(\cdot)$ denotes a kernel size (KS) 1×1 convolution layer with stride (S) 1 and C' output feature channels. In this paper, considering the interaction of pixels and smoothing effect of convolution layer, we designed pixel patch attention mechanism. Specifically, each query pixel patch tensor $Patch_Q \in \mathbb{R}^{patch_S\times patch_S\times C'}$ ($patch_S$ represents the patch size) times the transpose of corresponding key patch tensor $Patch_K^T \in \mathbb{R}^{C'\times patch_S\times patch_S}$ to form the pixel attention map. And $\mathcal{P}_\mathcal{V}$ is converted into a matrix $P_V \in \mathbb{R}^{C'\times HW}$. The output of pixel patch attention operator is computed as

$$Output_P = P_V \times \frac{1}{(patch_S)^2}(Patch_{Q_{ij}} \cdot Patch_{K_{ij}}^T), \tag{2}$$

where $i \in \mathbb{R}^H$ and $j \in \mathbb{R}^W$ that iterate the whole feature map and $Output_P \in \mathbb{R}^{C'\times HW}$ which is further converted back to $\mathbb{R}^{H\times W\times C'}$.

The choice of subsequent channel attention layer relies on the types of multi-scale attention block. For SA block, a $KS = 3$ and $S = 1$ convolutional layer is used to generate channel query $\mathcal{C}_\mathcal{Q}$. For DA block, a $KS = 3$ and $S = 2$ convolutional layer is employed to generate $\mathcal{C}_\mathcal{Q}$ and we apply a $KS = 3$ and $S = 2$ deconvolutional layer to generate $\mathcal{C}_\mathcal{Q}$ as

$$\mathcal{C}_\mathcal{Q} = G(Output_P) \in \mathbb{R}^{H_Q\times W_Q\times C''}, \tag{3}$$

where $G(\cdot)$ outputs C'' feature maps. Similar with the operation of $\mathcal{P}_\mathcal{K}$ and $\mathcal{P}_\mathcal{V}$, $\mathcal{C}_\mathcal{K}$ and $\mathcal{C}_\mathcal{V}$ are generated with a $KS = 1$ and $S = 1$ convolutional layer with C'' output feature channel. Then these three are converted to matrix $C_Q \in \mathbb{R}^{C''\times H_Q W_Q}$, channel key $C_K \in \mathbb{R}^{C''\times HW}$ and channel value $C_V \in \mathbb{R}^{C''\times HW}$. Finally, the output is computed as

$$Output = C_V \times Softmax(C_K^T C_Q), \tag{4}$$

and achieved with $Output \in \mathbb{R}^{H_Q\times W_Q\times C''}$. In conclusion, given $Input \in \mathbb{R}^{H\times W\times C}$ our proposed multi-scale attention

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.

block can generate feature maps $Output \in \mathbb{R}^{H_Q \times W_Q \times C''}$ with optional feature size and channel.

C. Memory Module With Hash Coding Addressing

As the Fig.2. shown, given the feature embedding vector z of the current image x as the input of memory module, it is first stored at memory slots. Simultaneously, it is mapped with hashing function to get the hash binary codes, which is stored at the hash table. Instantaneously, the similarity coefficients of input z and the whole hash table is computed with hamming distance. In the end, the corresponding embedding feature is taken from memory slots to feed the decoder. The proposed hash memory module involves a set of memory slots to preserve the encoding representations of normal instances and hash-based addressing for querying and retrieving. The memory slot matrix is denoted as $M = \{m_1, m_2, \ldots, m_i, \ldots, m_N\} \in \mathbb{R}^{N \times F}$, where N represents the number of slots and F means the dimension of each memorized item features. Each row vector m_i of M denotes a memory item. Theoretically, F is optional, and in practice, for reducing the computation cost, F is always set as same as the dimension of encoding representations. In the training phase, parameters of all memory slots are random initially. With the training on normal data iteratively, all slots are updated to preserve encoding representations of normal instances. Simultaneously, the memory module matrix M is mapped to a hash matrix $HS = \{h_1, h_2, \ldots, h_i, \ldots, h_N\} \in \{0, 1\}^{N \times F'}$. Specifically, given the memory module slot m_i, considering the following hash function:

$$r(m_i) = \underset{h \in \{0,1\}^F}{argmin} - f(m_i; \theta)^T)h, \tag{5}$$

the idea is to optimize the weights $f(\cdot; \theta) : m_i \rightarrow \mathbb{R}^{\mathbb{F}}$ that can activate the corresponding manifolds in the binary hash code h, consequently, hash the data M into a hash table HS that can denote the hash code as HS. We seek to optimize hash mapping function to hash similar instances into alike buckets, in other words, we wish to maintain the proximity of similar items and divided apart dissimilar items during the hash process [20]. Following the work [21], the hash function in Eq.(5) can be implemented with a single neural layer with trainable weights as follows:

$$h_i = \frac{1}{2}(sgn(fc(m_i) - 0.5\mathbf{1}) + 1), \tag{6}$$

where fc represents a fully connected layer that transform the memory feature vector with dimension of m_i to the dimension of the dimension of h_i, sgn denotes a sign activate function that can output one or negative one based on the input m_i, and $\mathbf{1}$ represents a ones vector with length h_i. Therefore, the binary codes h_i with the value one or zero is obtained.

Traditionally, given a query tensor from encoder network $z \in \mathbb{R}^{\mathbb{H} \times \mathbb{W} \times \mathbb{C}}$, [14] directly calculates the cosine similarity of z and each preserved item, which will be utilized as the attention coefficients for addressing. The attention addressing vector $w_{cos} \in \mathbb{R}^{\mathbb{N}}$ represents the similarity of the query z and each memory item m_i. Although it has advantages in quantification of low-dimension data, it cannot be a comprehensive and impactful evaluation standard of similarity of high-dimension and sparse data. In addition, w_{cos} is required for accessing N times memory module and meanwhile, it's observed from experiments that a larger enough N results in better work, so this method takes redundant cost. In this paper, the hamming distance HD

$$\begin{aligned} HD(h_z, h_i) &= \sum_{j=1}^{C} |(h_z)_j - (h_i)_j|, \\ (h_z)_j &= (h_i)_j \Rightarrow |(h_z)_j - (h_i)_j| = 0, \\ (h_z)_j &\neq (h_i)_j \Rightarrow |(h_z)_j - (h_i)_j| = 1, \end{aligned} \tag{7}$$

is introduced for computing the attention coefficients and similarity-based search on hash matrix HS. There are two major advantages of this mapping hash matrix HS: (i) greatly reducing the dimension of the calculative matrix from F (F equals to the feature maps' height H times feature maps' width W times the number of channel C) to F' (F' is set as 128 or 256 generally); (ii) the high efficiency of searching in the hash matrix via sorting hamming distance HD. The hamming distance weight vector $w_{HD} \in \mathbb{R}^{\mathbb{N}}$ represents the similarity of the query z and each memory slot. The smaller HD means the more homogeneous. Both soft and hard addressing methods can be conducted on our framework. The soft addressing takes the top smallest HD memory items and combines them with the coefficients that HD via a Softmax operation as follows:

$$\hat{h_z} = Softmax(w_{HD})M. \tag{8}$$

But the chance that the anomaly can be reconstructed well with a compound combination of memory normal instances cannot be ignored. The hard addressing is a more suitable method, as follows shown:

$$\hat{h_z} = argmin(w_{HD})M. \tag{9}$$

In Fig. 1 and Fig. 2, we provide the visualization of the hash memory module, which shows the output of hard addressing only retrieval the smallest hamming distance prototypical normal item. Then the latent representation $\hat{z}$ can be obtained via the hash corresponding memory matrix M. This method reduces the computation cost greatly and encourages the model to represent an item with fewer latent representation, which results in more informative features.

D. Encoder and Decoder

Encoder network is utilized for embedding input, which is encouraged to represent the input in a meaningful and informative manifold space. Given a sample x, the encoder maps it to an embedding representation z as follows:

$$z = f_e(x; \theta_e), \tag{10}$$

where θ_e denotes the parameters of the encoder network. Different from previous works [25], [26] that stack deep, linear CNNs to build networks, our encoder adopts the simple but effective proposed multi-scale attention block. As Fig. 2 shown, after resizing all inputs to size $H \times W \times C$, the encoder network established by three DA blocks and one SA block. Each DA block generates half-size feature maps. The dimension of final feature maps is $H/8 \times W/8 \times 512$. In addition,

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.

we proposed a semantic basis that generated by the SA block to represent the semantic information of encoding features, which is concatenated with decoding feature maps lately. Concretely, the semantic basis is plugged into decoder to resort to fully utilization of feature fusion. Our proposed UA and SA block is a simple yet effective heuristic-based method compared with most prevailing methods that simply sum up while fusing different feature sizes. Simultaneously, after getting retrieval $\hat{z}$ from our hash memory block as mentioned before, the decoder network consists of two SA blocks and three UA blocks. Each concatenated layer is passed to the SA block to reducing the channel. Then the UA block can generate new feature maps with double-size. The decoder is trained to reconstruct the input given a latent representation:

$$\hat{x} = f_e(\hat{z}; \theta_d). \tag{11}$$

Moreover, besides the $MSE = \|x - \hat{x}\|^2$ to measure the quality of reconstruction, we couple Earth-Mover distance [36] which is a criterion of data distribution as

$$EM(P_x, P_{\hat{x}}) = \inf_{\gamma \sim \prod(P_x, P_{\hat{x}})} \mathbb{E}_{(x,\hat{x})\sim\gamma}[\|x - \hat{x}\|] \tag{12}$$

between the input data distribution P_x and the generator data distribution $P_{\hat{x}}$. For joint probability distribution $\gamma \sim \prod(P_x, P_{\hat{x}})$, and for each marginal distribution is P_x or $P_{\hat{x}}$, the goal is to minimize the lower bound of the expected value of distance. With minimizing EM distance, it can let the generator data distribution be closer to the prior distribution. Moreover, it is easier to converge at the embedding space than using Kullback–Leibler (KL) divergence [36], which is common data distribution loss but it tends to infinity when there is no overlap between two distributions and has a mutation when states from non-overlap to overlap. With our loss function as

$$Loss = MSE(x, \hat{x}) + EM(x, \hat{x}), \tag{13}$$

the model can achieve better pixel reconstruction and realize more alike data distribution.

IV. Experiments

To demonstrate the robustness and generalization ability of our proposed MAMA Net, experiments are conducted on three datasets, including COVID-19 CT Images [27], COVID-19 X-Ray Images [37] and Reference Image Database to Evaluate Response (RIDER) Neuro MRI dataset [38] and evaluating the model with stratified k-fold cross-validation. Compared with various baseline models and state-of-the-art algorithms, the results prove the high effectiveness and excellent generalization of the proposed MAMA Net. Experiments are deployed on PyTorch [39] using Adam [40] optimizer with a learning rate of 0.01 with NVIDIA GeForce GTX 1080 graphics card for 70k iterations with mini-batch size of 8 samples. Code will be made available on https://github.com/DanielChen98/MAMA_NET_Pytorch.

A. Parameter Settings

As Fig. 4 shows, the training set only involves normal instances, which has no overlapping with the testing items. All anomalies are going with testing samples. In this experiments, we build the encoder network with three DA block ($KS = 3, S = 2$) and one SA block ($KS = 3, S = 1$). Firstly, the inputs are resized to 228×228. And taking the inputs, the first DA block with $C'' = 64$ generates $Conv\_1 \in \mathbb{R}^{114\times114\times64}$. Similarly, the second and third DA block outputs $Conv\_2 \in \mathbb{R}^{57\times57\times256}$ and $Conv\_3 \in \mathbb{R}^{29\times29\times512}$.

Simultaneously, the Semantic Basis $\in \mathbb{R}^{29\times29\times512}$ generated by SA block is used in our model and the encoding representation $z \in \mathbb{R}^F (F = 29 \times 29 \times 512)$ is flatten from $Conv\_3$. We implement the hash memory block with a set of memory slots $(N \times F)$ for recording and retrieving and a hash table $(N \times F')$. It's observed that a larger N has a better results and the effect of N will be discussed later. Briefly, the $N = 2000$ is the bottleneck of performance. The hash coding is achieve by Sigmoid activation function and its dimension F' is set as 128. The decoder network is composed of two UA block ($KS = 3, S = 2$) and two SA block ($KS = 3, S = 1$). UA block generates double-size feature maps and SA block is used to reduce channel: $Pre\_DeConv\_3 \in \mathbb{R}^{29\times29\times1024}$, $DeConv\_3 \in \mathbb{R}^{29\times29\times1024}$, $Pre\_DeConv\_2 \in \mathbb{R}^{57\times57\times512}$, $DeConv\_3 \in \mathbb{R}^{57\times57\times512}$, $DeConv\_2 \in \mathbb{R}^{114\times114\times64}$ and $Ouput(\hat{x}) \in \mathbb{R}^{228\times228\times3}$. The loss coupled with MSE and Wasserstein distance is applied on all experiments.

B. Experiments on COVID-19

We first conduct the experiments to detect the anomaly in COVID-19 CT [27] experiments, the training set only includes the non-COVID cases that contains 463 images with a composition of 36 images from LUNA, 195 from MedPix, 202 from PMC, and 30 from Radiopaedia. And the test set includes non-COVID samples that has168 CT images (164 of them from LUNA dataset and the rest 4 from Radiopaedia) and 100 COVID CT images from SIRM COVID-19 Database. And the experiments conducted on chest X-ray contains a total of 1808 images, which includes 225 COVID-19 chest X-ray images and 1583 normal images obtained from Cohen [37]. Similarly, the partition of normal samples to training set and test set is based on k-fold cross-validation strategy. In addition, all 225 COVID cases are consider as test set. The normal samples are divided into training set and testing set, following the setting used in [27], [37]. The memory size N of [27] is set as 2000 and [37] is 3000, which both are the tradeoff according its dataset as the ablation studies shown. The brief results of reconstruction of the normal and abnormal cases of MAMA Net are shown in Fig. 4. It's observed that the reconstruction of normal sample is almost same to the input, but the reconstruction image of COVID-19 case is distinct from the input and similar with the normal item which results in high reconstruction error.

The results are compared with traditional algorithm: one-class SVM (OC-SVM) [7], non-reconstruction methods based on deep learning: Visual Geometry Group (VGG16), ResNet, Dense-UNet [27], Inf-net [33], Zheng [32] and reconstruction methods: MemAE [14], Autoencoder-VGG16 (AE-VGG16), Autoencoder-ResNet101 (AE-ResNet101), TSC [2] and StackRNN [2]. Specifically, non-reconstruction methods are supervised learning that predicts the prediction of anomalies, and anomaly classification can be

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.

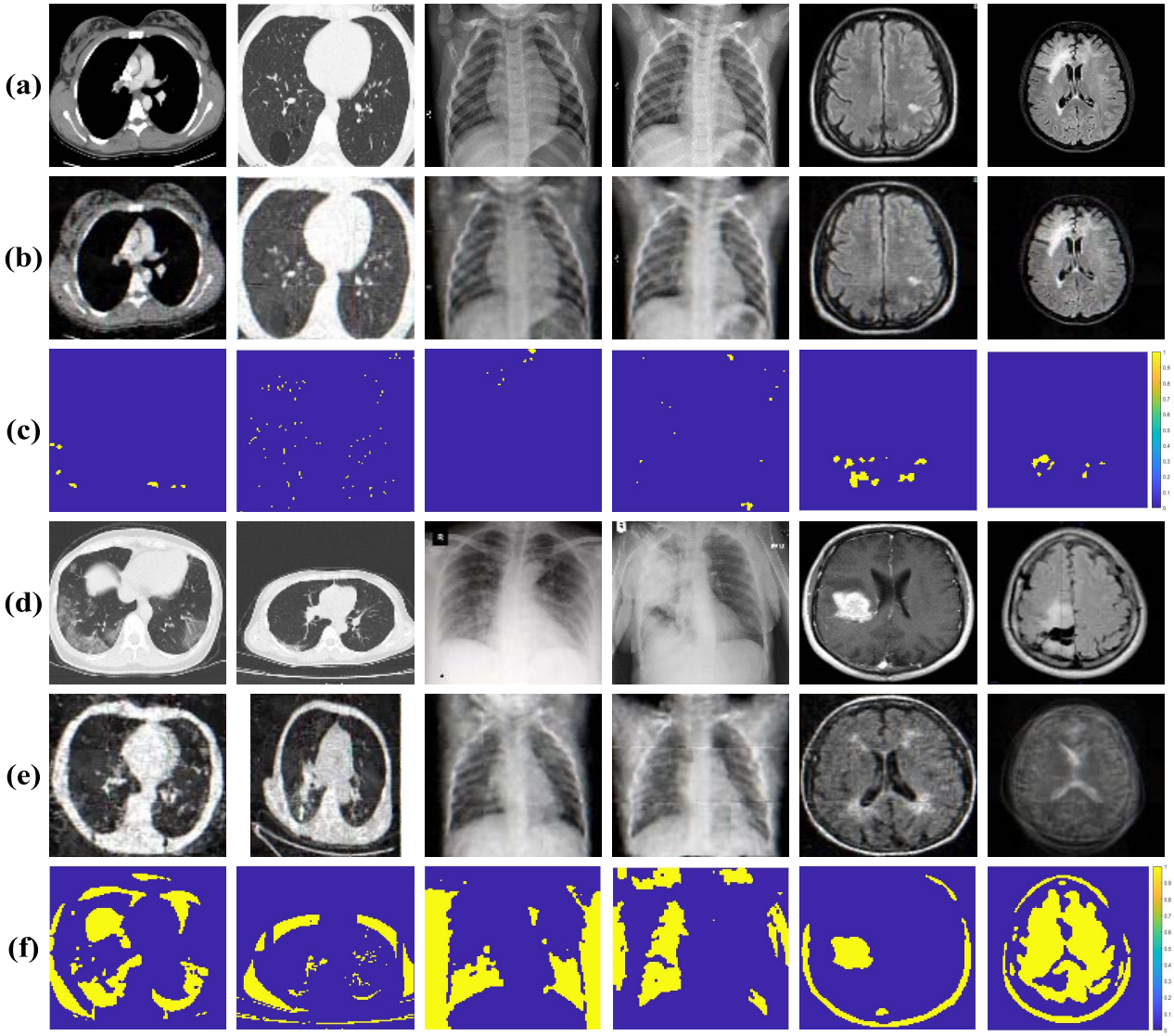

Fig. 4. The results of our proposed MAMA Net. Only need training on normal samples, our model can generate high reconstruction error of abnormal samples than normal samples.(a) testing inputs on normal samples; (b) reconstruction on normal samples; (c) reconstruction error of binary mask on normal samples; (d) testing inputs on anomaly; (e) reconstruction on anomaly; (f) reconstruction error of binary mask on anomaly.

evaluated by reconstruction error of reconstruction methods. AE-VGG16 and AE-ResNet101 denote the autoencoder network is built with VGG16 and ResNet101. In testing, we report the sensitivity, specificity, F1-score and Area Under Curve (AUC) as Table II shown, MAMA Net outperforms all reconstruction-based baselines and is more than a match for the state-of-the-art method which needs labels for training. We also do the Student's t-test [41] between results of MAMA Net and MemAE [14] that the p-value of all tests were performed at a significance level of $\alpha = 0.05$ (two-sided). The visualization and comparison of the reconstruction image of normal and COVID-19 cases are shown in Fig. 5, it's obvious that our MAMA Net has advantages in anomaly classification, feature extraction, and reconstruction quality.

C. Experiments on RIDER Neuro MRI

Furthermore, for proving the generalization of proposed MAMA Net, we conduct experiments on RIDER Neuro MRI [38] for evaluating the tumor detection. The training set consists of brain MRIs from 19 patients. T1 and T2-weighted MRIs are used, which contains a total of 349 MRIs, including 109 normal images and 240 abnormal images. Most baselines are introduced before, moreover, some former state-of-the-art methods [42], [43] are compared. The memory size N is also set as 2000. A short comparison of sensitivity, specificity, and F1-score is presented in Table II (all p-values tested between our model and [14] are less than 0.05), the obtained results prove the superiority of the proposed method in terms of brain tumor detection.

In the end, the normalized the reconstruction normality score p_u of u-th index image with range [0, 1] as follows

$$p_u = 1 - \frac{e_u - min_u(e_u)}{max_u(e_u) - min_u(e_u)}, \tag{14}$$

where e_u denotes the reconstruction error, between normal cases and abnormalities are compared with proposed MAMA Net as shown in Fig. 6, which provides that our proposed method an evident gap. And the comparison of training loss and speed are compared in Fig. 7, where our proposed model shows a stable loss and a high frames per second (FPS).

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.

TABLE II
SUMMARY OF TESTING SENSITIVITY, SPECIFICITY, F1-SCORE, AND AUC RESULTS ON COVID-19 CT [27], X-RAY IMAGES [37] AND RIDER NEURO MRI [38]

	Sensitivity			Specificity			F1-score			AUC		
	CT	X-Ray	RIDER MRI	CT	X-Ray	RIDER MRI	CT	X-Ray	RIDER MRI	CT	X-Ray	RIDER MRI
OC-SVM [7]	0.624	0.682	0.542	0.641	0.735	0.372	0.632	0.704	0.441	0.707	0.783	0.656
VGG16	0.787	0.830	0.691	0.736	0.804	0.660	0.761	0.817	0.675	0.802	0.834	0.711
ResNet101	0.794	0.847	0.656	0.831	0.813	0.794	0.812	0.829	0.718	0.836	0.857	0.785
Dense-UNet [28]	0.723	0.766	–	0.875	0.811	–	0.792	0.788	–	0.816	0.820	–
Inf-Net [34]	0.870	**0.894**	–	0.800	0.865	–	0.834	0.862	–	0.872	0.891	–
Zheng [33]	**0.907**	0.880	–	**0.911**	**0.946**	–	**0.909**	**0.912**	–	**0.932**	**0.935**	–
Pereira [44]	–	–	0.812	–	–	0.830	–	–	0.821	–	–	0.846
Abd-Ellah [45]	–	–	**0.818**	–	–	**0.846**	–	–	**0.832**	–	–	**0.861**
AE-VGG16	0.504	0.512	0.481	0.473	0.576	0.493	0.488	0.542	0.487	0.607	0.623	0.656
AE-ResNet101	0.623	0.605	0.660	0.601	0.644	0.683	0.616	0.624	0.671	0.668	0.712	0.707
TSC [2]	0.771	0.679	0.692	0.697	0.688	0.779	0.732	0.683	0.733	0.706	0.745	0.792
StackRNN [2]	0.712	0.833	0.846	0.821	0.810	0.819	0.763	0.821	0.832	0.810	0.843	0.831
MemAE [15]	0.855	0.871	0.832	0.842	0.858	**0.855**	0.848	0.864	0.843	0.924	0.911	0.858
MAMA Net (ours)	**0.901** ±0.06	**0.920** ±0.02	**0.873** ±0.03	**0.909** ±0.05	**0.938** ±0.04	**0.847** ±0.05	**0.905**	**0.929**	**0.859**	**0.957**	**0.969**	**0.883**

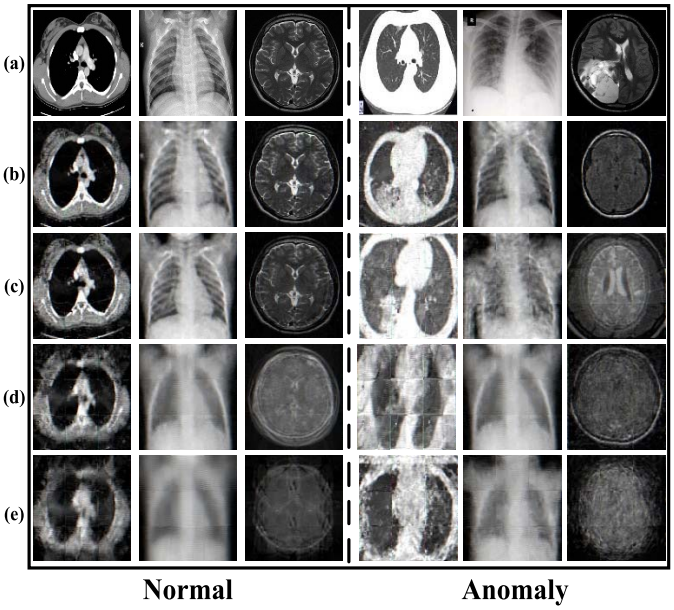

Fig. 5. The comparison of reconstruction normal cases (left), anomaly (right) of MAMA Net with other baselines. (a) input samples; (b) reconstruction of our MAMA Net; (c) reconstruction of MemAE [14]; (d) reconstruction of AE-ResNet101; (e) reconstruction of AE-VGG16. It can be seen that our proposed network has a similar reconstruction of normal samples and intrinsically different the reconstruction of anomalies, while other baselines cannot make a distinction between reconstruction of normal cases and anomaly.

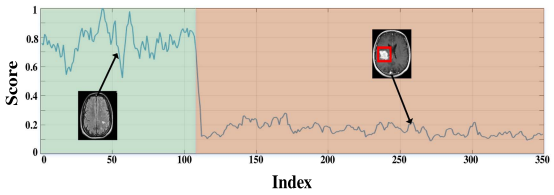

Fig. 6. Normality scores of RIDER Neuro MRI dataset. The green background represents the normal samples that hold high normality scores. And the score decreases immediately when anomalies appear.

D. Ablation Studies

As previous sections mentioned, comprehensive experiments comparisons have demonstrated the importance of the major components of the proposed MAMA Net, like hash-memory module and multi-scale attention block layer. In this section, we further conduct several ablation studies to analyze other different components more specifically.

1) Study of the Memory Size: We adopt all datasets to study the effect of the memory size N. The experiments with different N are shown in Fig. 8. The summary report of the F1-score is testing. In conclusion, with a large enough memory size (2000 as the red dash line), the model can robustly produce superior results with the datasets size from 100 images to 349 slices in our experiments.

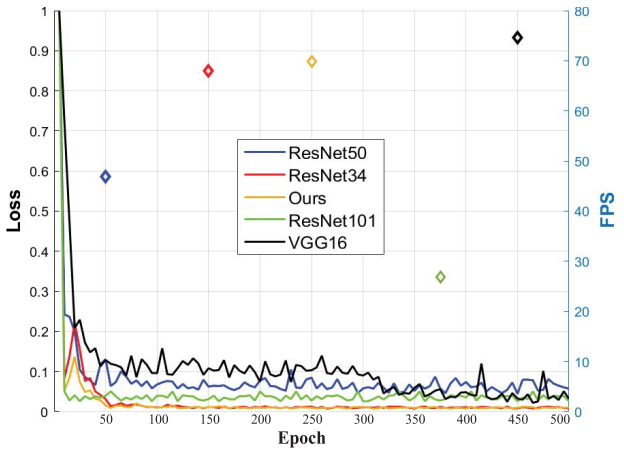

Fig. 7. The Loss of autoencoder network constructed with ResNet, VGG Net, and our proposed multi-scale attention block. The vanishing gradient that may stick in local optimization and unstable loss appears in VGG Net and ResNet 50/101. The network with only three ours blocks achieve minimum loss, which equals and perhaps surpasses ResNet34. The FPS also proves that our block with simple structure has advantages on detection speed.

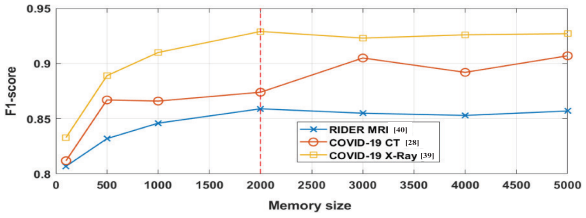

Fig. 8. Robustness to the setting of memory size. F1-score values of proposed MAMA Net with different memory size on COVID-19 CT, X-Ray Images and RIDER Neuro MRI are shown.

2) Study of the Semantic Basis: The deteriorated quality of decoder feature maps inherently restricts high-definition reconstruction, which makes normal samples and abnormalities without distinction. The semantic basis fully exploits the potential of encoding representations and the utilization of UA and SA block to fuse feature maps consistently achieve much better efficiency across a wide spectrum of resource. One brief comparison of utilization and discard of semantic basis

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.

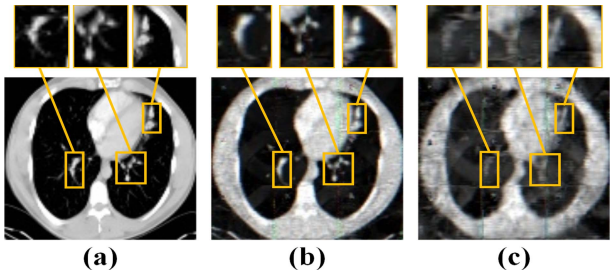

Fig. 9. The input (a) and the comparison of reconstruction image with utilization semantic basis (b) and without using semantic basis (c).

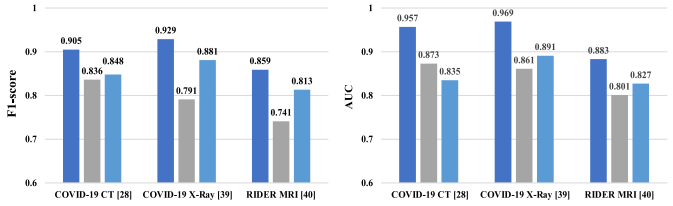

Fig. 10. Evaluation for the effectiveness of semantic basis and Wasserstein loss. Dark blue bar is the network with semantic basis and Wasserstein loss; Gray bar represents the model without semantic basis; Light blue bar is the network without Wasserstein loss.

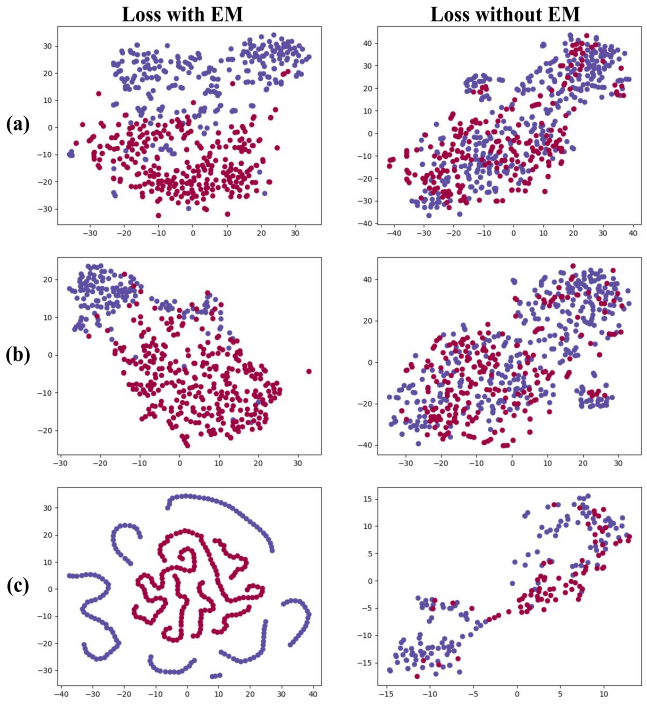

Fig. 11. The t-SNE visualization of embedded features (points) z. The comparison of with EM (left) and without EM (right) on (a) COVID-19 CT Images (b) COVID-19 X-Ray Images and (c) RIDER Neuro MRI.

is shown in Fig. 9 that it is obvious that the reconstruction keeps high fidelity with utilization semantic basis. In the end, we report the evaluation for the effectiveness of semantic basis as shown in Fig. 10. The performance remarkably increase at introducing semantic basis. In our model, the semantic basis is adopted as Fig. 2. shown.

3) Study of the Wasserstein Loss: The Wasserstein (EM) distance not only can measure two distributions similarity, also it provides every transferability of probability density. Here, we compare the features obtained from the hidden layer with EM and without EM of a trained MAMA Net on all datasets we used as Fig. 11 shown. It's obvious that with introducing EM distance into loss function, the robustness of embedding representation improved and the domain shift is mitigated. The validity of Wasserstein loss is also proven in performance as Fig. 10 shown.

V. Discussion and Conclusion

In this paper, we proposed a hash addressing memory autoencoder with the multi-scale attention block to detect the anomaly, especially for COVID-19 detection. With the reconstruction of the input by our encoder and decoder network, the anomaly produce high reconstruction error while the normal samples generate a lower error. Multi-scale attention block is designed to mitigates nowadays challenges of restricted stationary convolution operators with combining pixel patch attention and channel attention layer, which is conveniently plugged into any network for sampling, downsampling, and upsampling. Due to the memory is trained only to record the prototypical normal cases, it can reconstruct the normal samples well and amplify the reconstruction error of the anomalies [14]. But the soft-addressing via cosine similarity looks like a fish out of water, and a hash memory module is proposed for fast retrieving. To the best of our knowledge, this is the first time to introduce the hash addressing memory module into autoencoder. In addition, with coupling mean square error with Wasserstein distance over input to reconstruction data, the network resorts to robust data distribution. Experiments on various datasets prove the effectiveness and generalization of MAMA Net. Our proposed module achieves better performance than other baselines, while our model is a more general framework that can be flexibly applied to various types. In the future, we will apply our model on more challenging datasets, investigate quicker and easier memory module and addressing methods. Furthermore, designing taxonomy anomaly detection system is our next objective.

References

[1] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel, "Multiple-instance learning for anomaly detection in digital mammography," *IEEE Trans. Med. Imag.*, vol. 35, no. 7, pp. 1604–1614, Jul. 2016.
[2] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 341–349.
[3] J. Sharpnack, A. Rinaldo, and A. Singh, "Detecting anomalous activity on networks with the graph Fourier scan statistic," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 364–379, Jan. 2016.
[4] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, "A novel coronavirus outbreak of global health concern," *Lancet*, vol. 395, no. 10223, pp. 470–473, Feb. 2020.
[5] H. Kang *et al.*, "Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2606–2614, Aug. 2020, doi: 10.1109/TMI.2020.2992546.
[6] F. Y. Edgeworth, "On discordant observations," *Phil. Mag.*, vol. 23, no. 5, pp. 364–375, 1887.
[7] Y. Chen, X. Sean Zhou, and T. S. Huang, "One-class SVM for learning in image retrieval," in *Proc. Int. Conf. Image Process.*, 2001, pp. 34–37.
[8] P. Perera and V. M. Patel, "Learning deep features for one-class classification," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5450–5463, Nov. 2019, doi: 10.1109/TIP.2019.2917862.
[9] C. C. Tan, *Autoencoder Neural Networks: A Performance Study Based On Image Reconstruction, Recognition And Compression*. New York, NY, USA: Academic, 2009.

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.

[10] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception by hierarchical Bayesian models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[11] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 481–490.

[12] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognit.*, vol. 58, pp. 121–134, Oct. 2016.

[13] V. Vercruyssen, W. Meert, G. Verbruggen, K. Maes, R. Baumer, and J. Davis, "Semi-supervised anomaly detection with an application to water analytics," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 527–536.

[14] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[15] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[16] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly Detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[17] D. Amso and G. Scerif, "The attentive brain: Insights from developmental cognitive neuroscience," *Nature Rev. Neurosci.*, vol. 16, no. 10, pp. 606–619, Oct. 2015.

[18] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10076–10085.

[19] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. VLDB*, 1999, pp. 518–529.

[20] D. Wu, Q. Dai, J. Liu, B. Li, and W. Wang, "Deep incremental hashing network for efficient image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9061–9069.

[21] V. Gattupalli, Y. Zhuo, and B. Li, "Weakly supervised deep image hashing through tag embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10367–10376.

[22] M. S. Nixion and A. S. Aguado, *Feature Extraction & Image Processing for Computer Vision*. London, U.K.: Academic, 2012.

[23] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2921–2928.

[24] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[25] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[26] M. Jonathan, M. Ueli, and C. Dan, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Artif. Neural Netw. Mach. Learn.*, 2011, pp. 52–59.

[27] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-dataset: A CT scan dataset about COVID-19," 2020, *arXiv:2003.13865*. [Online]. Available: http://arxiv.org/abs/2003.13865

[28] X. Wang *et al.*, "A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2615–2625, Aug. 2020, doi: 10.1109/TMI.2020.2995965.

[29] O. Gozes *et al.*, "Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis," 2020, *arXiv:2003.05037*. [Online]. Available: https://arxiv.org/abs/2003.05037

[30] S. Wang *et al.*, "A deep learning algorithm using CT images to screen for corona virus disease (COVID-19)," *medRxiv*, Jan. 2020, doi: 10.1101/2020.02.14.20023028.

[31] X. Ouyang *et al.*, "Dual-sampling attention network for diagnosis of covid-19 from community acquired pneumonia," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2595–2605, Aug. 2020.

[32] Z. Chuansheng *et al.*, "Deep learning-based detection for COVID-19 from chest CT using weak label," *medRxiv*, May 2020, doi: 10.1101/2020.03.12.20027185.

[33] D. Fan *et al.*, "Inf-Net: Automatic COVID-19 lung infection segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, Aug. 2020.

[34] C. Li, J. Zhu, and B. Zhang, "Learning to generate with memory," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1177–1186.

[35] W. Lu *et al.*, "Unsupervised sequential outlier detection with deep architectures," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4321–4330, Sep. 2017, doi: 10.1109/TIP.2017.2713048.

[36] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, 2017, pp. 214–223.

[37] J. Paul Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," 2020, *arXiv:2003.11597*. [Online]. Available: http://arxiv.org/abs/2003.11597

[38] *The Cancer Imaging Archive, RIDER NEURO MRI Database*. Accessed: Sep. 14, 2018. [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/RIDER+NEURO+MRI

[39] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 8026–8037.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[41] Seize and Theresa, "Studentês t-test," *Southern Med. J.*, vol. 70, p. 1299, Dec. 1977.

[42] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240–1251, May 2016.

[43] M. K. Abd-Ellah, A. I. Awad, A. A. M. Khalaf, and H. F. A. Hamed, "Two-phase multi-model automatic brain tumour diagnosis system from magnetic resonance images using convolutional neural networks," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, p. 97, Dec. 2018.

Authorized licensed use limited to: CHANGSHA UNIV OF SCIENCE AND TECHNOLOGY. Downloaded on September 17,2021 at 04:20:22 UTC from IEEE Xplore. Restrictions apply.