DefQ: Defensive Quantization against Inference Slow-down Attack for Edge Computing

Han Qiu, Tianwei Zhang, Tianzhu Zhang, Hongyu Li, and Meikang Qiu *Senior Member, IEEE*

***Abstract*—The novel multi-exit Deep Neural Network (DNN) architectures provide a new optimization solution for efficient model inference in edge systems. Inference of most samples can be completed within the first few layers on an edge device without the need to transmit them to a remote server. This can significantly increase the inference speed and system throughput, which is particularly beneficial to the resource-constrained scenarios. Unfortunately, researchers proposed an inference slow-down attack against this technique, where an external adversary can add imperceptible perturbations on clean samples to invalidate the multi-exit mechanism. In this paper, we propose a Defensive Quantization (`DefQ`) method as the *first* defense against the inference slow-down attack. It is designed to be lightweight and can be easily implemented in off-the-shelf camera sensors. Particularly, `DefQ` introduces a novel quantization operation to preprocess the input images. It is capable of removing the perturbations from the malicious samples and preserving the correct inference exit points and prediction accuracy. Meanwhile, it has little impact on the clean samples. Extensive evaluations show that `DefQ` can effectively defeat the inference slow-down attack and well protect the efficiency of edge systems.**

***Index Terms*—Edge Computing, Deep Learning, Security, Inference Slow-down.**

I. Introduction

In the past decade, researchers have proposed various Deep Learning (DL) algorithms and models (e.g., Convolutional Neural Networks (CNNs) [1], Recurrent Neural Networks (RNNs) [2], Deep Reinforcement Learning (DRL) [3]), to solve complex tasks in different domains, such as Computer Vision, Natural Language Processing (NLP), and Autonomous Driving. Those state-of-the-art models have been extensively commercialized in many products, and perfectly integrated with modern edge computing systems. By hosting Deep Neural Network (DNN) models on the edge devices, the Internet of Everything (IoE) system becomes more intelligent to interpret and interact with the physical world in a more efficient fashion.

New emerging DNN models are becoming more complicated, with an increased number of parameters, layers, and computations. This guarantees they have higher performance and generalization to handle different types of data, at the cost of powerful computing capabilities. In general, the required computation for DL research doubles every few months, resulting in an increase of about 300,000× from 2012 to 2018[1]. This phenomenon adds difficulties to deploy state-of-the-art models on the IoT devices, which are generally resource- and computation-constrained. This becomes more severe in some critical scenarios which have high requirements for the inference speed and throughput [4]. Novel solutions are urgently needed to optimize the utilization of DNN models on tiny computing devices and small-scale systems.

In order to deploy increasingly complex DNN models into the edge computing scenarios such as Intelligent IoT, many different methodologies have been proposed to simplify the computation of the DNN inference process or to compress the DNN-related storage. They can be classified into the following categories. First, more compact DNN architectures were designed to adapt to the edge and mobile devices, such as MobileNets [5], SqueezeNet [6] and ShuffleNets [7]. They can achieve comparable accuracy with those complicated models, using much fewer parameters and computations. Besides, Adaptive Neural Networks (AdNN) [8] was proposed aiming at saving the energy consumption of inference by dynamically deactivating parts of its model based on the need of the inputs. However, this technique cannot reduce the model size, hindering the deployment of large models on small devices. Second, researchers proposed model compression [9] to reduce the model size while preserving the performance. Different methods were introduced to achieve this goal, including model pruning [10], [9], quantization [11], precision reduction [12], distillation [13], etc. According to [14], a typical ResNet-50 model can be compressed by removing 75% of its parameters while preserving similar performance. There still exists a trade-off between the compression ratio and model accuracy drop. Third, split learning was proposed for distributed inference [15], [16], [17], [18], [19]. Instead of compressing and putting the entire DNN model to one edge device, they aim to fragment the DNN model into multiple parts by the layer, and distribute them to different devices. A common practice is to deploy the first few layers on the resource-constrained edge device while offloading the rest layers to the remote server for acceleration. This collaborative mode can remarkably enhance the inference speed and throughput.

Recently, researchers proposed the multi-exit DNN architecture [20], which allows the model to make predictions for certain input samples at earlier stages. When a sample can be classified by a front layer with high confidence, the model can simply finish the inference task and output the results,

Han Qiu is with Institute for Network Sciences and Cyberspace, BNRist, Tsinghua University, Beijing 100084, China. Email: qiuhan@tsinghua.edu.cn.

Tianwei Zhang is with Nanyang Technological University, Singapore, 639798. Email: tianwei.zhang@ntu.edu.sg.

Tianzhu Zhang is with Nokia Bell Labs, Nozay, France, 91620. Email: tianzhu.zhang@nokia-bell-labs.com.

Hongyu Li is with Beijing University of Posts and Telecommunications, Beijing, China, 100876. Email: l543306408@bupt.edu.cn.

Meikang Qiu (corresponding author) is with Texas A&M University, Texas, USA, 77843. Email: qiumeikang@yahoo.com.

[1] openai.com/blog/ai-and-compute

without involving the rest layers. This early-exit mechanism can significantly save inference time and energy. It can be adopted cooperatively with the split learning: by deploying different layers (exit points) to the edge device and remote server, prediction of most samples can be completed at the local side, avoiding the time cost from network transmission and remote server computation. This is much more efficient compared to the original edge-cloud systems.

Unfortunately, the multi-exit DNN architecture is vulnerable to the inference slow-down attack, which can invalidate the early-exit mechanism, and force the inference samples to go through the entire model. Hong et al. [21] designed DeepSloth attack to achieve this goal. The adversary can use the gradient-based technique to generate imperceptible perturbations and add them to the clean samples. Then, these malicious samples will confuse each internal classifier for prediction, and fail to meet the early-stop criteria at each point. Thus, they have to travel across each layer on the edge device and remote server, as well as the network, to obtain the final results. Experiments in [21] demonstrated that the DeepSloth attack can achieve 100% success rate against the existing multi-exit models. Moreover, the model accuracy is also terribly compromised. Although this attack can bring severe threats to the edge systems in terms of efficiency and performance, to the best of our knowledge, there are currently no defense solutions to address this issue yet.

In this paper, we propose a Defensive Quantization (`DefQ`), the *first* methodology to defeat the inference slow-down attack in the edge computing scenario. The essential component of `DefQ` is a lightweight transformation function to process all the input samples before feeding them into the multi-exit DNN models. Particularly, we use a statistical manner to understand the malicious perturbations of the inference slow-down attack in the frequency domain and design a defensive quantization table for the preprocessing operation. During inference, the transformation function adopts this defensive quantization table to remove the potential malicious perturbations in the frequency domain. Meanwhile, it will not affect the normal samples. We conducted extensive evaluations on three mainstream multi-exit models and three l bound-based DeepSloth attacks. Experimental results show that `DefQ` can effectively mitigate the inference slow-down threat, and recover the early-exit mechanism for compromised samples. Besides, `DefQ` can also improve the prediction accuracy of these samples.

The key contributions of this paper are:

- the first defense method against the inference slow-down attack in intelligent edge computing systems
- a statistical method to understand the malicious perturbations brought by the inference slow-down attack
- an effective yet lightweight defensive quantization method to mitigate the inference slow-down attack and preserve the model's accuracy.

This paper is organized as follows. Section II discusses the research background and related works. Section III presents the threat model and defense requirements. Section IV analyzes the attack and describes the defense details. Section V presents the evaluation results. We discuss and list future works in Section VI and conclude in Section VII.

II. Background and Related Works

In this section, we briefly present the background and relevant works about intelligent edge systems, multi-exit DNN models, and inference slow-down attacks.

A. Artificial Intelligence on the Edge

The development of DL technology makes today's IoT ecosystems more intelligent and powerful. It becomes practical to deploy DNN models on edge devices for complex tasks. A conventional Artificial Intelligence of Things (AIoT) system consists of multiple frontend sensors for data collection and preprocessing, backend edge devices, and remote servers [22] for DNN inference. They are normally connected via wireless channels (e.g., 5G, Wi-Fi) to collaboratively sense, interpret and understand the environment. Such intelligent IoT systems have been widely adopted in many scenarios, such as face authentication [23], fire alarming system [24], and remote monitoring [25].

Modern edge systems are required to process the sensor data promptly. For instance, there can be a large number of high-resolution cameras generating real-time data continuously for online analysis [26]. However, the processing speed is restricted by the limited computing capabilities and resources of edge devices. One promising solution is to leverage the powerful compute servers to accelerate the inference process. We can split a DNN model into two parts, and offload the second part to the remote server. The collaborative model across the edge and server can increase the inference speed even when we consider the network latency. The design of multi-exit DNN models can better support such model split and enhance the edge system efficiency, as introduced below.

B. Multi-exit DNN Architecture

To accelerate the inference on resource-constrained devices, one promising research direction is to build multi-exit DNN architectures, which can selectively make predictions at earlier layers [27]. The key idea is to insert multiple classifiers (i.e., exit points) at different layers in a DNN model. The sample to be analyzed will go through each layer one by one, and each classifier attempts to make a prediction. If one classifier has sufficient confidence to predict this sample, i.e., the stop criteria are met, the inference task for this sample will be done at this exit point, and the predicted label will be assigned. Figure 1 shows the structure of a multi-exit DNN model. A number of works have proposed different algorithms and methods to support this feature. For instance, Huang et al. [20] designed the Multi-Scale Dense Networks (MSDNets) architecture, which can provide any time predictions and avoid any loss of prediction accuracy when directly modifying existing DNN architectures. [28] proposed the Shallow-Deep Network (SDN), which can directly convert a well-trained DNN model into a multi-exit one by inserting internal classifiers. The basic structure of these internal classifiers are designed as a feature reduction step followed by a fully connected layer classifiers.

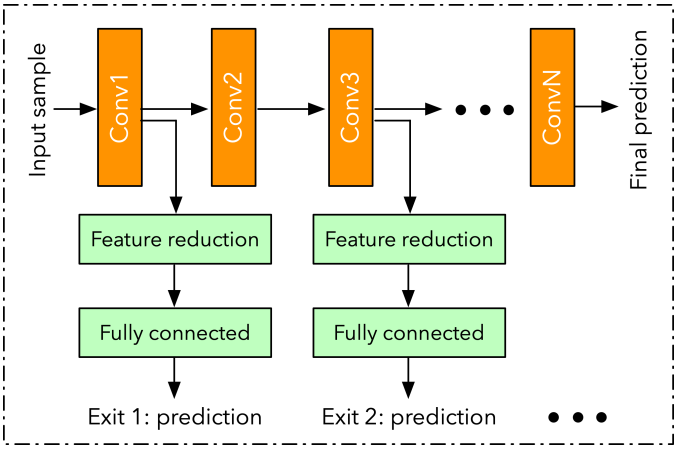

Figure 1. Building a Shallow-Deep Network (SDN) [28] multi-exit model by adding multiple exit points for prediction on a typical CNN model.

The inference will stop and the sample will exit once the sample's result at one exit point meets a pre-set threshold.

This multi-exit DNN architecture provides great opportunities to enable distributed inference with higher efficiency on the edge computing system. For instance, we can deploy the first few layers and exit points of any DNN model on the local edge device, and migrate the rest computation to the remote powerful server. Then, some of the input samples can be predicted and exit only on the edge device without any requirement for transmission and further computation cost on remote servers. Hence, even a tiny edge device can be involved for the complex model serving tasks since inference of a certain ratio of samples can be completed within the first few samples on the edge device to reduce the network latency. This is much more efficient than the traditional DNN split learning task [16], [17], [18], [19], where every sample has to go through all the layers on the edge and server to finish the prediction.

C. Inference Slow-down Attack

Inference Slow-down attack is proposed [21], aiming to slow down the inference speed of multi-exit models. An adversary can inject imperceptible perturbation on a clean sample, which makes the internal classifier fail to meet the early-exit threshold at each exit point. Then this sample has to go through every layer and obtain the final prediction results only at the last layer. This can significantly increase the inference time and computation energy. Such perturbation can be generated using the gradient-based approach to disable all the exit points. Note that the goal of this attack is different from Adversarial Examples (AEs), which aim to change the prediction label.

This inference slow-down attack is particularly severe for the edge computing context. First, it significantly compromises the inference efficiency by forcing all the samples to pass all the convolutional layers, resulting in a huge waste of time and energy for the multi-exit model inference. Second, it causes a majority of samples to be transmitted from the edge to the cloud for the final prediction. The incurred network latency can amplify the inference time by 1.5∼5×, negating the benefits of split learning for the edge-cloud setting. Third, as a side effect, the model accuracy will also be decreased by the added perturbation. To the best of our knowledge, there is currently a lack of defense solutions for this severe threat.

III. Threat Model and Defense Requirements

Threat Model. We consider an Artificial Intelligent edge system for computer vision tasks (e.g., image classification, object detection, etc). This system consists of three entities: frontend sensors keep collecting the images at a high sampling rate, and sending them to edge devices and remote servers, which collaboratively host a multi-exit DNN model for inference. Since an edge device has limited computing capabilities and memory, it only deploys the first few layers and exit points, while offloading the rest computations to the remote server. We focus on computer vision applications for two reasons. First, vision sensors like cameras are one of the most widely-used IoT devices in our daily life. Computer vision tasks are also commonly adopted in many scenarios, e.g., video surveillance [29], face authentication [23], autonomous driving [30], etc. Second, compared to other sensor data, vision sensors can produce a larger volume of real-time streaming data with higher throughput. Thus, edge systems for computer vision are in more urgent need of efficient inference solutions.

The entire edge system (e.g., frontend sensors, edge devices, remote servers, and their communication networks) is assumed to be trusted. So we do not consider the security threats from inside the system (e.g., botnet Mirai [31], Hajime [32]) or Man-in-the-Middle network attacks. The adversary is outside of the edge system, attempting to spoof the sensor data by adding malicious perturbations on the physical objects, camera lens, or digital images. He has full knowledge of the target DNN model but is not allowed to tamper with the parameters (e.g., DNN backdoor attacks [33]). The adversary's goal is to *feed malicious images to the system to significantly reduce its efficiency: the DNN model has to go through all the layers to process these images with much longer time cost, resulting in more energy consumption and less throughput.* Such inference slow-down attacks have been introduced in [21]. Attacks aiming at other purposes such as AEs, backdoor, and data poisoning are not within the scope of this paper.

Defense Requirements. The goal of this paper is to design a novel and effective methodology for mitigating inference slow-down attacks against multi-exit DNN models. We assume the defender will not modify or enhance the target model itself. For instance, the DNN model can be purchased from a model vendor and deployed with specific settings. It may not be allowed or practical for users to alter or retrain the model. The defender only implements transformation functions in front of the DNN model to preprocess the input images. The functions need to meet the following requirements.

- *Effectiveness*: they should be able to rectify the malicious samples, forcing them to follow the correct exits in the model. Meanwhile, they should not affect the exit points of benign samples.
- *Accuracy-preserving*: they should have little impact on the prediction accuracy of malicious or benign samples.

- *Lightweight*: the computation of these functions should not be too heavy to affect the operation of the edge systems, considering the limited onboard computing capabilities and resources of edge devices.

IV. Proposed Defense Methodology

We present a novel approach `DefQ` to protect multi-exit DNN models from inference slow-down attacks. First, we give a detailed analysis on a representative attack – DeepSloth [21] in Section IV-A. Then, we provide the methodology overview in Section IV-B, followed by the detailed designs in Section IV-C and defense security analysis in IV-D.

A. DeepSloth Attack Analysis

Hong et al. proposed the DeepSloth attack [21], which can slow down the inference process of a multi-exit model by adding malicious perturbations to the benign samples. The core idea of the DeepSloth attack is to make the samples never meet the exit thresholds at each exit point of the DNN model. Specifically, by adding carefully crafted perturbations on input samples, an adversary can manipulate the representation of each layer in the model and push the outputs of the internal classifier at each exit point towards a uniform distribution. Apparently, a uniform distribution of the internal classifiers' outputs cannot meet any thresholds to exit at these exit points. Hence, each internal classifier cannot make confident decisions about the samples, and deeper layers are needed for classification. Figure 2 shows the visual results of an attack example on a ResNet-56-SDN model with 27 exit points. A clean sample (a) will exit at the 1st exit point. DeepSloth generates imperceptible perturbations bounded by the L_1-norm (b), L_2-norm (c), or L_∞-norm (d), which make the samples miss all the 27 exit points and obtain the prediction at the final layer. Note that this attack is fundamentally different from the previous adversarial attacks, which aim to change the predictions and trigger the misclassification at earlier layers.

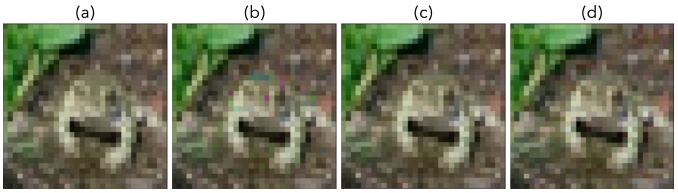

Figure 2. An example of DeepSloth attack on a ResNet-56 model: (a) a clean sample exits at the 1st exit point; while (b) L_1-based, (c) L_2-based, and (d) L_∞-based attack samples miss all the 27 exit points.

Efficiency analysis. The primary target of the DeepSloth attack is inference efficiency. The adversary's goal is to cause the inference samples to fail the criteria of all the early exit points, until reaching the final prediction layer (Figure 1). This will compromise the system efficiency from two aspects: for each infected sample, the inference time will be significantly delayed as it needs to go through more convolutional layers or even the network between the edge device and remote server. For the entire system, the edge device needs to spend more time to classify the samples, hence the inference throughput is reduced and more energy is wasted.

Accuracy analysis. Preserving or decreasing the prediction accuracy is not a target of DeepSloth. Although the authors in [21] claimed that they try not to affect the prediction results, our experimental results indicate the prediction accuracy is still decreased to some extent. How to design more efficient inference slow-down attacks without affecting the model accuracy is beyond the scope of this work.

B. Overview of `DefQ`

Figure 3 shows an overview of our proposed methodology, `DefQ`. The core of `DefQ` is a defensive quantization-based function deployed in the frontend sensors. It preprocesses each captured image before sending it to the edge server. It will not affect the execution of benign images, while effectively removing the malicious perturbations added by the inference slow-down attack. This quantization step can be integrated with the image compression and formatting operations to reduce the size of transmitted images.

In a benign case, the analysis of most images will be completed in the first few layers residing on the edge device, without going to the remote server. However, the malicious samples will go through more layers across the edge device and remote server with longer execution time and network latency. Our quantization function aims to remove the perturbations from these malicious samples and make their exit on the edge device again. Besides, as DeepSloth attack can also affect the prediction results of malicious samples, our quantization function can even improve the model accuracy over these samples.

In order to design a qualified defensive quantization function, we generate the quantization table by statistically computing the influence on the frequency domain in DeepSloth. We aim to use a small set of malicious samples generated from DeepSloth to get a statistical frequency influence, which can inspire us to further design the quantization table (More details are described in Section IV-C).

C. Defensive Quantization

The key idea of `DefQ` is to utilize a defensive quantization operation to remove the perturbations in the frequency domain in a lightweight manner. For an input image, we first transform it from the spatial domain to the frequency domain based on a Discrete Cosine Transform (DCT) [34]. This DCT coefficient distribution represents the energy contribution of each frequency band. Previous works [35], [36] have shown that different frequency bands have different influence levels on DNN learning and inference.

Here, we adopt a 2D-DCT which first cut images into grids with a setting size. Pixels in each grid are transformed into the frequency space via 2D-DCT as shown in Eq. (1): $f(x, y)$ refers to the input image and C is the coefficients for the DCT calculation. The values of C can then be calculated according to the grid size $N \times N$ (in this paper, we set $N = 8$).

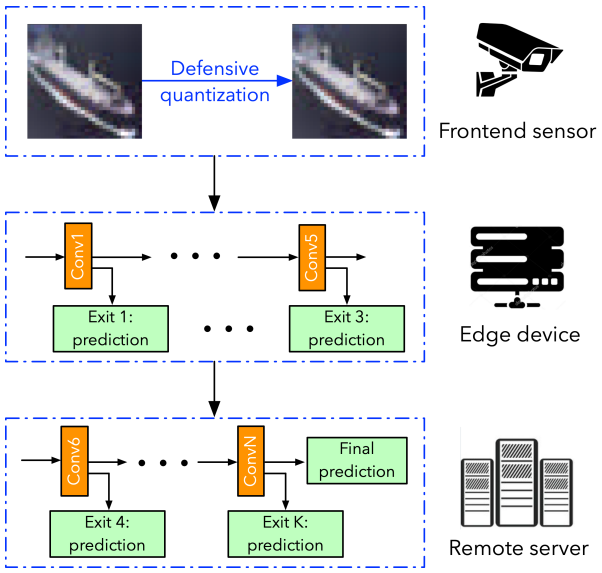

Figure 3. Overview of our defensive quantization-based methodology.

$$F(u,v) = \frac{2}{N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} C_{x,y}(u,v) f(x,y)$$

$$C_{x,y}(u,v) = \alpha(u)\alpha(v)\cos\left[\frac{\pi(2x+1)u}{2N}\right]\cos\left[\frac{\pi(2x+1)v}{2N}\right]$$

$$\alpha(x) = \begin{cases} \sqrt{\frac{1}{N}}, & \text{when } x = 0 \\ \sqrt{\frac{2}{N}}, & \text{otherwise} \end{cases} \quad (1)$$

After the 2D-DCT, we design a specific quantization method by analyzing the modification in the frequency domain caused by the DeepSloth attack. Based on the adversary's viewpoint, to bypass a quantization-based defense, the malicious perturbation in terms of pixel values must be large enough to introduce significant modifications in the frequency domain to further influence the quantized results. Therefore, the quantization-based defense can make it significantly difficult or even impossible for the adversary to craft the perturbations. Previous works aim to design a specific quantization table to mitigate the adversarial attacks for misleading the DNN prediction [35]. For the DeepSloth attack, we find the malicious modifications in the frequency domain have special patterns that can be used to design a defensive quantization table.

Our novel method to generate the quantization table is presented in Algorithm 1. We generate the table Q_{def} in a statistical learning manner by summarizing the frequency patterns of the malicious samples. Here we first build a clean sample set I_M by collecting M clean samples from the dataset ($\mathbb{R}$) with width H, height W, and 3-layer of colors. Then, we use the DeepSloth L_∞ attack to generate a malicious sample set $\widehat{I}_M$ from the I_M. Therefore, $\widehat{I}_M$ contains M images that miss all exit points of the model. These steps will build the input of our algorithm as $I_M \in \mathbb{R}^{M\times H\times W\times 3}$ and $\widehat{I}_M \in \mathbb{R}^{M\times H\times W\times 3}$. Note that sing other perturbation bounds (L_1, or L_2) for malicious sample generation will lead to similar results. Our algorithm first fragments all the images into 8×8 blocks (Lines 3-4) on all three color layers. Then, these blocks in the spatial domain (I in Line 6) are collected from all the images' color channels for both the benign image set and malicious set ($\widehat{I}$ in Line 7). By conducting DCT-2D on all the 8×8 blocks, we compare the difference of DCT frequency coefficients (Line 8) to understand the perturbation ratio in the frequency domain brought by DeepSloth attack. This is calculated by comparing the difference of all the blocks in the frequency coefficients from the clean and malicious sets accordingly. By calculating the average modification ratio, we get the modification levels for different frequency bands (Line 13). Then, we normalize these statistical results of the modification in the frequency space, and remap them back into a range between 20 and 60 (Line 14) as our quantization table Q_{def}. We also try different ranges such as [20, 80] or [20, 100], which yield similar defense results (see the ablation study in Section V-D).

ALGORITHM 1: Generating the quantization table Q_{def}.

Input: clean set $I_M \in \mathbb{R}^{M\times H\times W\times 3}$,
malicious set $\widehat{I}_M \in \mathbb{R}^{M\times H\times W\times 3}$,
a zero matrix of size 8×8 Q_0
Output: defensive quantization table Q_{def}

1 **for** i *in* $1 \sim M$ **do**
2 **for** $I_{i,channel}$ *in* I_i **do**
3 $n_w = W/8$, $n_h = H/8$;
4 $\mathcal{G}(I_i) = \{(x_m, y_n) | (m,n) \in (0 \sim n_w, 0 \sim n_h)\}$;
5 **for** (x_m, y_n) *in* $\mathcal{G}(I_i)$ **do**
6 $F(I_{i,channel}) = DCT(I_{i,channel}(x_{m-1} : x_m, y_{n-1} : y_n))$;
7 $F(\widehat{I}_{i,channel}) = DCT(\widehat{I}_{i,channel}(x_{m-1} : x_m, y_{n-1} : y_n))$;
8 $Dif(i)_{m,n} = (|F(I_i)| + 1)/(\left|F(\widehat{I}_i)\right| + 1)$;
  $Q_0 = Q_0 + Dif_{m,n}$;
9 **end**
10 **end**
11 **end**
12 $Q_0 = Average(Q_0, M \times n_w \times n_h \times 3)$;
13 $Q_{def} = (Q_0 - min(Q_0)/max(Q_0)) \times 40 + 20$;
14 **return** Q_{def};

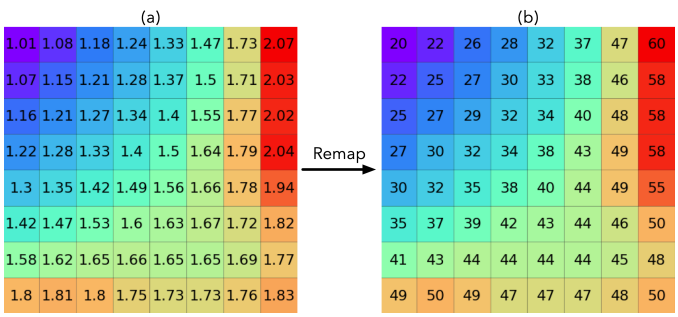

Figure 4. DCT frequency space statistical results of DeepSloth attack for random 1000 samples (a) and our defensive quantization table Q_{def} (b).

The statistical results in the frequency domain of comparing the clean and malicious samples are given in Figure 4(a). We observe that the low-frequency bands including the DC value are slightly changed. The high-frequency bands (the lower

right corner) are modified by the DeepSloth attack, but the middle frequency bands (the upper right corner) are changed the most. Figure 4(b) shows our quantization table Q_{def} after remapping the statistical results.

D. Security Analysis

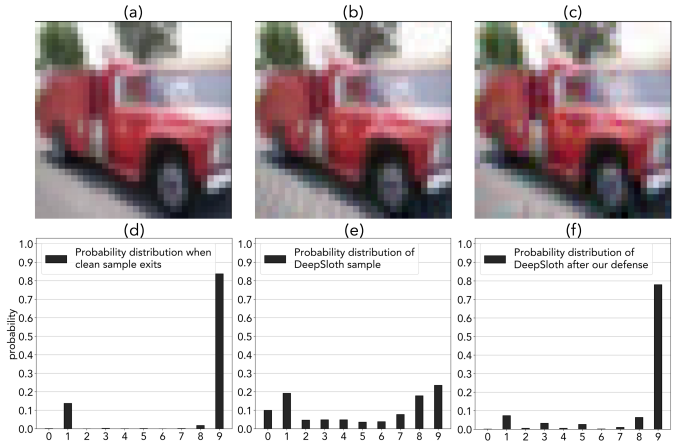

Figure 5. An example of our defense on DeepSloth attack. The first row shows a clean image (a), a malicious image before without our defense (b), and with our quantization preprocess (c). The second row shows the corresponding confidence scores from the classifier at the 2nd exit point. Images (a) and (c) quit at the 2nd exit point, while image (b) quit at the final layer.

We use a simple example to demonstrate the effects of the feature modification by our defensive quantization. Figure 5 shows a clean image (a), malicious DeepSloth image (b), and the output of our quantization function for this malicious image (c). The second row shows the predicted confidence scores at the 2nd exit point corresponding to the images in the first row. We observe that for the clean image, the classifier has high confidence to predict it as a "truck" (d). This meets the early-exit criteria and the task will be completed with the correct label. However, for the malicious image without our defense, the probability distribution of the confidence score is rather even, and the classifier cannot make the decision (e). It has to send the image to the subsequent layers for further analysis, and finally, the task is done at the last layer. Now with our defensive quantization method, the malicious perturbation in the frequency space is removed. The confidence score is recovered to the correct one, and the task can be done at the 2nd exit point. This indicates the DeepSloth attack is successfully defeated, and the multi-exit scheme works again.

V. Evaluations

We comprehensively evaluate our proposed methodology over three mainstream models. We first evaluate the defense effectiveness of `DefQ` on DeepSloth attacks with different techniques (L_1, L_2, and L_∞). Second, we measure the impact of `DefQ` on the model accuracy to validate its accuracy-preserving property. Considering the defense requirements in Section III, we evaluate and give the results about the *effectiveness* in Section V-B, *accuracy-preserving* and *lightweight* in Section V-C, respectively.

A. Experimental Configuration

We consider the image classification task on the CIFAR-10 dataset. It contains 50,000 images for training and 10,000 images for testing. Each image has a size of $32 \times 32 \times 3$ and belongs to one of ten classes. All pixel values are normalized within the range of $[0, 1]$. We consider three state-of-the-art DNN models (ResNet-56 [37], VGG-16, and MobileNet [28]), and use the SDN structure [21] to convert them into multi-exit models. The corresponding SDN-ResNet-56, SDN-VGG-16 and SDN-MobileNet models have 27, 14 and 14 exit points, respectively. To reproduce the DeepSloth attack, we follow the configurations in [21] under a white-box setting (i.e., the adversary has full knowledge of the victim SDN model) and generate the malicious samples with the L_1, L_2, and L_∞ based techniques. We use Pytorch 1.8 [38] as the backend for model training and inference. All the experiments are run on a server with an Intel (R) Core (TM) i9-10900K CPU@3.70 GHz and two Nvidia GeForce RTX 3090 GPUs.

B. Defense Effectiveness Evaluation

We adopt two metrics to quantify the effectiveness of `DefQ`. The first one is the Top-3 Exit Ratio (T3-ER), which measures the percentage of inference samples that can be completed within the first three exit points deployed on the edge device (Figure 3). A higher T3-ER indicates higher inference efficiency. The second metric is Efficacy (EFCY) from [21]. This metric is defined to quantify a model's ability to utilize its exit points for inference. It is a normalized value between 0 and 1: a value closer to 1 indicates more input samples will exit earlier to save inference time. We consider different thresholds for early-exit in SDN-based models. At each exit point, a confidence score will be selected and compared with a threshold to determine whether this sample should take this exit point. A smaller threshold will lead more samples to exit earlier, but the accuracy may be decreased. We set two thresholds to make the Relative Accuracy Drop (RAD) is within 5% and 10% of its maximum accuracy, respectively.

We randomly select 1000 clean samples from the testing dataset and generate the malicious samples. Table I presents the T3-ER values of different samples without and with `DefQ`. First, we observe that about 55%, 20%, and 53% clean samples can meet the threshold of RAD=5% within the first three exit points for the three models respectively. The percentage is even higher when we set the threshold as RAD=10%. This can significantly reduce the computation cost and edge-server transmission cost. The introduced quantization function has little impact on the clean samples. In the SDN-MobileNet model, `DefQ` can even slightly improve T3-ER. Second, we observe that with the DeepSloth attack, almost zero samples will exit within the first three exit points under the two thresholds. Most samples must be sent from the edge device to the remote server with higher latency. With `DefQ`, we see that T3-ER is significantly improved, indicating the successful mitigation of the DeepSloth attack.

Table II shows the EFCY results which are similar with T3-ER. This confirms the effectiveness of `DefQ` on both benign and malicious samples against different techniques.

Table I
TOP-3 EXIT RATIO OF SAMPLES INCLUDING CLEAN SAMPLE, DEEPSLOTH SAMPLES OF DIFFERENT CONSTRAINS WITH OR WITHOUT DEFQ.

RAD	Sample	SDN-VGG-16		SDN-ResNet-56		SDN-MobileNet	
		original	DefQ	original	DefQ	original	DefQ
5%	Clean	0.55	0.54	0.20	0.21	0.53	0.57
	DeepSloth (L_∞)	0.00	0.26	0.00	0.07	0.00	0.27
	DeepSloth (L_2)	0.35	0.45	0.09	0.14	0.36	0.44
	DeepSloth (L_1)	0.27	0.34	0.08	0.11	0.27	0.30
10%	Clean	0.70	0.69	0.34	0.34	0.77	0.81
	DeepSloth (L_∞)	0.00	0.41	0.00	0.12	0.05	0.48
	DeepSloth (L_2)	0.50	0.58	0.21	0.26	0.57	0.69
	DeepSloth (L_1)	0.40	0.56	0.17	0.21	0.45	0.61

Table II
EFCY OF SAMPLES INCLUDING CLEAN SAMPLE, DEEPSLOTH SAMPLES OF DIFFERENT CONSTRAINS WITH OR WITHOUT DEFQ.

RAD	Sample	SDN-VGG-16		SDN-ResNet-56		SDN-MobileNet	
		original	DefQ	original	DefQ	original	DefQ
5%	Clean	0.78	0.76	0.56	0.54	0.82	0.81
	DeepSloth (L_∞)	0.02	0.53	0.00	0.44	0.01	0.55
	DeepSloth (L_2)	0.42	0.66	0.16	0.44	0.43	0.70
	DeepSloth (L_1)	0.31	0.52	0.12	0.30	0.31	0.59
10%	Clean	0.85	0.84	0.69	0.67	0.92	0.92
	DeepSloth (L_∞)	0.06	0.66	0.01	0.41	0.06	0.76
	DeepSloth (L_2)	0.57	0.78	0.27	0.58	0.65	0.85
	DeepSloth (L_1)	0.45	0.76	0.21	0.55	0.54	0.79

We also analyze the distribution of exit points over 1000 samples. We use the SDN-VGG-16 model as an example, which contains 14 exit points and one final prediction point. Figure 6 shows the results for the clean samples (a), malicious samples (b), and transformed malicious samples with our solution (c). We find that most clean samples can complete the inference process successfully within the first 10 exit points. In contrast, the majority of DeepSloth samples will be forced to take the final point. With our defense, the DeepSloth samples become normal again after the transformation, and more than 85% of them can be done within the first 10 exit points. This proves that our solution can maintain the functionality of the multi-exit mechanism under the inference slow-down attacks.

C. Model Accuracy Evaluation and Lightweight Discussion

As discussed in Section III, another defense requirement is to preserve the model accuracy on the inference samples. Table III shows the average accuracy of clean samples and DeepSloth samples generated by different techniques without and with DefQ. (1) We find that the introduced quantization transformation has a small influence on the accuracy of clean samples. (2) The DeepSloth attack can significantly decrease the accuracy of malicious samples, in addition to the exit points. (3) Our DefQ can effectively remove the perturbations and increase the model accuracy of those samples, which is close to that of clean samples. This proves our defense method achieves the *accuracy-preserving* requirement.

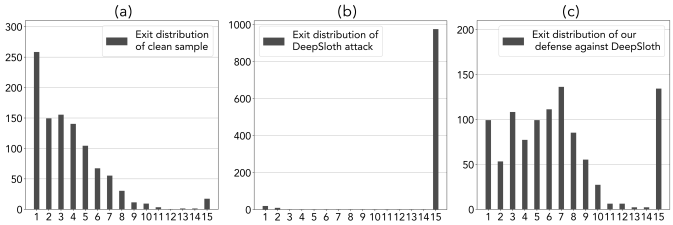

Figure 6. Exit distribution comparison between (a) 1000 clean samples, (b) clean samples attacked by DeepSloth, and (c) our defense against DeepSloth on SDN-VGG-16 model with RAD = 5.

DefQ is designed as a special quantization operation after the DCT-2D transformation. Today, most frontend camera sensors have the computing capabilities of compressing images before sending them to edge devices. The most common image compression procedure such as the JPEG standard includes a typical DCT-2D transformation, quantization, and coding operations. Therefore, our defensive quantization function can be easily integrated into this pipeline without any additional computation requirements. Therefore, DefQ meets the *lightweight* requirement.

Table III
MODEL ACCURACY OF DIFFERENT SAMPLES.

RAD	Sample	SDN-VGG-16		SDN-ResNet-56		SDN-MobileNet	
		original	DefQ	original	DefQ	original	DefQ
5%	Clean	0.84	0.82	0.80	0.77	0.78	0.76
	DeepSloth (L_∞)	0.13	0.72	0.21	0.69	0.21	0.67
	DeepSloth (L_2)	0.67	0.79	0.57	0.73	0.61	0.74
	DeepSloth (L_1)	0.52	0.71	0.45	0.71	0.54	0.70
10%	Clean	0.80	0.78	0.75	0.72	0.73	0.72
	DeepSloth (L_∞)	0.16	0.70	0.21	0.67	0.24	0.66
	DeepSloth (L_2)	0.68	0.75	0.58	0.70	0.62	0.71
	DeepSloth (L_1)	0.57	0.69	0.48	0.70	0.59	0.66

D. Ablation Study of the Quantization Table

In Section IV-C, we statistically calculate the influence brought by the DeepSloth attack in the frequency domain and remap them into an integer range to generate the quantization table. We compare different ranges of the quantization table during the remapping: [20, 60], [20, 80], and [20, 100]. We adopt the SDN-VGG-16 model and the L_∞ attack, and the evaluation results are shown in Table IV. We observe that the effectiveness and accuracy metrics are similar for different quantization tables. This demonstrates that DefQ is effective based on the statistical understanding of the influence in the frequency domain.

Table IV
ABLATION STUDY ON THE RANGE OF THE QUANTIZATION TABLE.

RAD	Sample		SDN-VGG-16		
			T3-ER	EFCY	ACC
5%	Clean		0.55	0.78	0.84
	DeepSloth (L_∞)	Original	0.00	0.02	0.13
		DefQ ([20, 60])	0.26	0.53	0.72
		DefQ ([20, 80])	0.25	0.54	0.72
		DefQ ([20, 100])	0.27	0.55	0.70
10%	Clean		0.70	0.85	0.79
	DeepSloth (L_∞)	Original	0.00	0.06	0.16
		DefQ ([20, 60])	0.41	0.66	0.70
		DefQ ([20, 80])	0.38	0.63	0.72
		DefQ ([20, 100])	0.42	0.65	0.71

VI. DISCUSSION AND FUTURE WORK

Advanced attacks and defenses. In this paper, we mainly focus on the mitigation of DeepSloth, the most common inference slow-down attack against computer vision tasks. Even the

adversary knows our quantization function, he cannot generate the desired samples using the gradient-based technique, since the quantization process is non-differentiable. One possible attack method is to adopt the Backward Pass Differentiable Approximation (BPDA) technique to approximate the gradient of the quantization step. However, BPDA adversarial examples can also be mitigated by advanced gradient obfuscation methods [39], and we believe they can defeat such advanced inference slow-down attacks as well. In the future, we will design and evaluate more sophisticated inference slow-down attacks and countermeasures.

Future testing in real-world systems. We measure the impacts of attacks and our defense on three popular models. We plan to implement our `DefQ` in real-world edge systems as the future work. Since we have analyzed the discussed the practical computation steps and costs in Section V-C so our lightweight purpose will hold in real-world edge systems. For the future implementation, we will adopt the edge devices (e.g., raspberry pi, Nvidia Jetson Nano) and remote cloud services to host the computer vision tasks (e.g., face authentication [23], remote monitoring [25]). We will also generate the DeepSloth samples to attack the system, and use `DefQ` to mitigate them. The inference speed, network latency cost and energy consumption will be measured to demonstrate its practicality and effectiveness.

VII. Conclusion

In this paper, we propose `DefQ`, a novel approach to effectively mitigate the inference slow-down attacks. We design a quantization-based transformation to preprocess input images, which is able to remove the perturbations on the malicious samples. Then these samples will follow the correct exit points to improve the inference efficiency and system throughput. Meanwhile, their prediction accuracy is also significantly increased. We also validate that `DefQ` is able to maintain the efficiency and accuracy of clean samples. `DefQ` is lightweight and can be deployed in off-the-shelf image sensors to protect the computer vision tasks in Artificial Intelligent edge systems.

Acknowledgement

This work is supported by the Natural Science Foundation of China under Grant No. 62106127.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[4] K. Gai and M. Qiu, "Optimal resource allocation using reinforcement learning for iot content-centric services," *Applied Soft Computing*, vol. 70, pp. 12–21, 2018.

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[6] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[7] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[8] X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez, "Skipnet: Learning dynamic routing in convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 409–424.

[9] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[10] C. Dai, X. Liu, H. Cheng, L. T. Yang, and M. J. Deen, "Compressing deep model with pruning and tucker decomposition for smart embedded systems," *IEEE Internet of Things Journal*, 2021.

[11] S. Xu, H. Li, B. Zhuang, J. Liu, J. Cao, C. Liang, and M. Tan, "Generative low-bitwidth data free quantization," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–17.

[12] Z. Cai, X. He, J. Sun, and N. Vasconcelos, "Deep learning with low precision by half-wave gaussian quantization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5918–5926.

[13] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," *arXiv preprint arXiv:1802.05668*, 2018.

[14] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," *arXiv preprint arXiv:1710.09282*, 2017.

[15] J. Hauswald, T. Manville, Q. Zheng, R. Dreslinski, C. Chakrabarti, and T. Mudge, "A hybrid approach to offloading mobile image classification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 8375–8379.

[16] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," *ACM SIGARCH Computer Architecture News*, vol. 45, no. 1, pp. 615–629, 2017.

[17] A. E. Eshratifar, M. S. Abrishami, and M. Pedram, "Jointdnn: An efficient training and inference engine for intelligent mobile cloud computing services," *IEEE Transactions on Mobile Computing*, 2019.

[18] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, pp. 148–162.

[19] ——, "Attacking and protecting data privacy in edge–cloud collaborative inference systems," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9706–9716, 2020.

[20] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," *ICLR*, 2018.

[21] S. Hong, Y. Kaya, I.-V. Modoranu, and T. Dumitraş, "A panda? no, it's a sloth: Slowdown attacks on adaptive multi-exit neural network inference," *ICLR*, 2020.

[22] M. Qiu, Z. Ming, J. Wang, L. T. Yang, and Y. Xiang, "Enabling cloud computing in emergency management systems," *IEEE Cloud Computing*, vol. 1, no. 4, pp. 60–67, 2014.

[23] Y. Mao, S. Yi, Q. Li, J. Feng, F. Xu, and S. Zhong, "A privacy-preserving deep learning approach for face recognition with edge computing," in *USENIX Workshop Hot Topics Edge Computing*, 2018, pp. 1–6.

[24] Z. Shouran, A. Ashari, and T. Priyambodo, "Internet of things (IoT) of smart home: privacy and security," *International Journal of Computer Applications*, vol. 182, no. 39, pp. 3–8, 2019.

[25] G. Chen, T. X. Han, Z. He, R. Kays, and T. Forrester, "Deep convolutional neural network based species recognition for wild animal monitoring," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 858–862.

[26] Q. Zhang, T. Huang, Y. Zhu, and M. Qiu, "A case study of sensor data collection and analysis in smart city: provenance in smart food supply chain," *International Journal of Distributed Sensor Networks*, vol. 9, no. 11, p. 382132, 2013.

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[28] Y. Kaya, S. Hong, and T. Dumitras, "Shallow-deep networks: Understanding and mitigating network overthinking," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3301–3310.

[29] Y. Tang, C. Zhang, R. Gu, P. Li, and B. Yang, "Vehicle detection and recognition for intelligent traffic surveillance system," *Multimedia tools and applications*, vol. 76, no. 4, pp. 5817–5832, 2017.
[30] B. Wu, F. Iandola, P. H. Jin, and K. Keutzer, "Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 129–137.
[31] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "DDoS in the IoT: Mirai and other botnets," *Computer*, vol. 50, no. 7, pp. 80–84, 2017.
[32] S. Herwig, K. Harvey, G. Hughey, R. Roberts, and D. Levin, "Measurement and analysis of Hajime, a peer-to-peer IoT botnet." in *NDSS*, 2019.
[33] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2019, pp. 707–723.
[34] R. Reininger and J. Gibson, "Distributions of the two-dimensional DCT coefficients for images," *IEEE Transactions on Communications*, vol. 31, no. 6, pp. 835–839, 1983.
[35] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: DNN-oriented JPEG compression against adversarial examples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 860–868.
[36] H. Qiu, Q. Zheng, T. Zhang, M. Qiu, G. Memmi, and J. Lu, "Toward secure and efficient deep learning inference in dependable IoT systems," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3180–3188, 2020.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.
[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.
[39] H. Qiu, Y. Zeng, Q. Zheng, S. Guo, T. Zhang, and H. Li, "An efficient preprocessing-based approach to mitigate advanced adversarial attacks," *IEEE Transactions on Computers*, 2021.

Han Qiu received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2011, the M.S. degree from Telecom-ParisTech (Institute Eurecom), Biot, France, in 2013, and the Ph.D. degree in computer science from the Department of Networks and Computer Science, Telecom-ParisTech, Paris, France, in 2017. He worked as a postdoc and a research engineer with Telecom Paris and LINCS Lab from 2017 to 2020. Currently, he is an assistant professor at Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China. His research interests include AI security, big data security, and the security of intelligent transportation systems.

Tianwei Zhang is an assistant professor in School of Computer Science and Engineering, at Nanyang Technological University. His research focuses on computer system security. He is particularly interested in security threats and defenses in machine learning systems, autonomous systems, computer architecture, and distributed systems. He received his Bachelor's degree at Peking University in 2011, and the Ph.D. degree at Princeton University in 2017.

Tianzhu Zhang is a research engineer at Nokia Bell Labs. He received his B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2012. Afterward, he received the M.S. degree in 2014, and the Ph.D. degree in 2017, both from Politecnico di Torino, Turin, Italy. From 2017 to 2019, he was a PostDoc researcher at Telecom ParisTech and LINCS, under a research grant from Cisco Systems. He joined Nokia Bell Labs in August 2020. His current research interests include SDN/NFV, Artificial Intelligence, Edge Computing, Robotics, Big Data, and log analysis.

Hongyu Li received his B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2021. He is currently pursuing his master's degree at the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China. His research interest includes deep learning and AI security.

Meikang Qiu received the BE and ME degrees from Shanghai Jiao Tong University and received Ph.D. degree of Computer Science from University of Texas at Dallas. He is the Department Head and tenured full professor of Texas A&M University Commerce. He is ACM Distinguished Member and IEEE Senior Member. He is the Chair of IEEE Smart Computing Technical Committee. He has published 20+ books, 600+ peer-reviewed journal and conference papers, including 100+ IEEE/ACM Transactions papers. His research interests include Cyber Security, Big Data Analysis, Cloud Computing, Smarting Computing, Intelligent Data, Embedded systems, etc. He is an Associate Editor of 10+ international journals, including IEEE Transactions on Computers, IEEE Transactions on Cloud Computing, IEEE Transactions on Big Data, and IEEE Transactions on SMC. He is ACM Distinguished Member (2019), Highly Cited Researcher (2020, Web of Science), and IEEE Distinguished Visitor (2021-2023).