A Radio-Frequency-Based 2-D Convolutional Layer using Transmissive Intelligent Surfaces

Jingyuan Zhang, Haige Chen, and Douglas M. Blough
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, United States
Emails: jingyuan@ece.gatech.edu, hchen425@gatech.edu, doug.blough@ece.gatech.edu

2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall) | 979-8-3315-1778-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/VTC2024-Fall63153.2024.10757747

***Abstract*—A novel convolutional layer based on transmissive intelligent surfaces (TISs), which operates in the radio-frequency (RF) domain, is introduced for analog over-the-air (OTA) computation in this paper. To be specific, each RF convolutional layer comprises three TISs placed sequentially to perform a 2-D convolution operation. A method for designing TIS transmission coefficients, TIS locations, and TIS element spacing is proposed to execute the 2-D convolution of $I * K = O$. The transmission coefficients of the second TIS encapsulate information about the kernel K, and the output (O) of the third TIS is the convolutional result of K and the input signal (I), which is the input to the first TIS. To validate the proposed design, a simple neural network featuring a single convolutional layer with one kernel is tested. The simulation results demonstrate that, with a practical size of the proposed design and adequate signal power transmitted to the TISs, the neural network incorporating the proposed TIS-based convolutional layer achieves a good approximation of performance compared to a neural network with the classic complex-valued convolutional layer. This validates the feasibility of the proposed design and the potential for offloading convolution operations from digital processors to the RF domain.**

***Index Terms*—over-the-air, analog computation, transmissive intelligent surface, 2-D convolutional layer, neural networks**

I. Introduction

With the development of machine learning (ML) and the proliferation of Internet-of-Things (IoT) devices, the demand for ML processing in edge applications is increasing [1], [2]. Edge AI, which refers to implementing ML models on edge devices including robots, IoT devices, and cellphones, has attracted substantial research interest due to the advantages of enhanced privacy, real-time processing, and lower dependence on Internet connectivity.

Due to the widespread deployment of wireless networks including 5G, Wi-Fi, ultra-wideband (UWB), Bluetooth, etc, IoT devices process a lot of radio-frequency (RF) data for communication purposes. Furthermore, wireless sensing has emerged as a technique to enhance sensing capability of IoT devices. This technique relies on extracting target information from variations in wireless signals reflected from the target, where many applications, such as gesture and movement recognition, rely on ML for precise detection [3], [4]. However, ML model execution on edge devices faces challenges including limited memory and computational capability.

Recently, the concept of over-the-air (OTA) computation has been proposed to offload computation from digital processors into the RF domain [5]–[7]. The key idea of OTA computation is to directly manipulate the RF signals that carry sensing or communication information in the RF domain, such that the signal propagation emulates specific mathematical operations. One method of OTA computation, which is explored in this paper, is to process RF signals using intelligent surfaces. An intelligent surface consists of an array of low-power, low-cost electromagnetic elements, each of which can be independently configured to manipulate incoming RF signals [8], [9]. These surfaces can be fabricated as either intelligent reflecting surfaces (IRSs) or transmissive intelligent surfaces (TISs). With IRSs, signals can only be reflected back off the surface whereas with TISs, signals can either be reflected off or transmitted through the surface, thereby providing 360° of coverage for the outgoing RF signals [7], [8], [10]. Intelligent surfaces enable OTA computation by adjusting the transmission or reflection coefficients to control incoming RF signals and produce desired outgoing RF signals. OTA computation using intelligent surfaces has the potential to reduce the memory and computational demands on edge devices by offloading computations to separate devices operating in the RF domain.

OTA computation using intelligent surfaces has the following benefits: (1) the weights of ML models can be embedded in pre-designed transmission/reflection coefficients of intelligent surface elements, potentially reducing memory requirements on edge devices; (2) computation can be offloaded to the RF domain, as OTA computation occurs when incoming RF waves interact with intelligent surface elements; and (3) intelligent surfaces offer power and cost advantages, as they are composed of passive electromagnetic elements.

There are a few studies exploring OTA computation for neural networks using IRSs, albeit research on this topic remains limited. In [6], a 1-D convolutional layer in the RF domain based on IRSs is introduced. The design in [6] uses N transmitters and N IRSs to implement an N-tap finite impulse response (FIR) filter. However, the proposed system requires tight synchronization and alignment among multiple transmitters, and higher power consumption and latency compared to its digital equivalent on a graphics processing unit (GPU) or field programmable gate array (FPGA), as reported in [6]. Furthermore, this method is limited to 1-D convolution, and the exploration of RF-based 2-D convolution remains unaddressed. Given that 2-D convolution is a fundamental operation in widely used convolutional neural networks (CNNs), this motivates us to explore new ways of achieving OTA computation that can extend to 2-D convolution while,

at the same time, reducing the need for tight synchronization.

In this paper, a 2-D convolutional layer in the RF domain using multiple TISs placed in close proximity is proposed. TISs, rather than IRSs, are employed here since transmissive arrays are more suitable for consecutive placement, allowing for sequential RF signal manipulation. Moreover, multiple IRSs placed in proximity will cause unwanted interference due to reflections among IRSs, which makes mathematical analysis of signals to achieve a specific processing goal intractable. To be specific, three TISs are placed sequentially to emulate a 2-D convolution operation ($I * K = O$), where I is the input signal of the first TIS, O is the output signal of the third TIS, and the second TIS contains the information of the kernel K. In this paper, we propose and validate the design of 2-D convolution based on three TISs through theoretical analysis and simulation. The key insight is that by applying the Fourier transform to the input signal I and kernel K, the convolution operation turns into matrix multiplication operations that resemble RF signal transmission through several TISs. Although there are limitations regarding power consumption of the proposed structure and the acquisition of channel information between adjacent TISs, which are discussed in detail in Sec. V, this study represents the first attempt, to the best of the authors' knowledge, to explore 2-D convolution in the RF domain. Sec. V also discusses open problems and practical challenges related to real-world OTA implementation of 2-D convolution.

The rest of the paper is organized as follows. Sec. II introduces the system model and the overall structure of the proposed design. Sec. III discusses the design details. Sec. IV contains simulation results to validate and evaluate the proposed design. In Sec. V, the takeaways and limitations of the design are discussed. Finally, Sec. VI concludes.

II. System Model

A new convolutional layer design using three TISs is presented for implementing 2-D convolution of $I*K = O$ with RF signals. In this section, the system model and a general overview of the proposed design are introduced.

A. System Model

As shown in Fig. 1, the convolutional layer contains three square TISs sequentially placed along the z-axis, with the TISs' centers aligned along this axis. Without loss of generality, it is assumed that the center of TIS 1 is placed at location $[0, 0, 0]$. It is also assumed that TIS i has $N_i \times N_i$ elements where $i \in \{1, 2, 3\}$, and the size of each TIS element is $d_e \times d_e$. Let D_i be the distance between the centers of TIS i and TIS $i + 1$ where $i \in \{1, 2\}$, and d_i be the element spacing of TIS i. Moreover, let $\mathbf{m} = [m_x, m_y] \in \mathbb{R}^{1\times 2}$ denote the index of an element on an TIS array, where $m_x, m_y \in \{-\frac{N-1}{2}, -\frac{N-1}{2}+1, -\frac{N-1}{2}+2, \cdots, \frac{N-1}{2}-2, \frac{N-1}{2}-1, \frac{N-1}{2}\}$ if the square TIS has $N \times N$ elements.

B. Transmission Model

The model of signal transmission between two consecutive TISs is introduced in this section. Let $h^i_{\mathbf{m},\mathbf{n}}$ represent the channel between element $\mathbf{m}$ on TIS i and element $\mathbf{n}$ on TIS $i + 1$. According to [11], $h^i_{\mathbf{m},\mathbf{n}}$ can be modeled based on Rayleigh-Sommerfeld diffraction as follows:

$$h^i_{\mathbf{m},\mathbf{n}} = A_t \cos(\phi^i_{\mathbf{m},\mathbf{n}})\left(\frac{1}{2\pi d^i_{\mathbf{m},\mathbf{n}}} - \frac{j}{\lambda}\right)\frac{e^{j\frac{2\pi d^i_{\mathbf{m},\mathbf{n}}}{\lambda}}}{d^i_{\mathbf{m},\mathbf{n}}}, \quad (1)$$

where $A_t = d_e^2$ is the size of one TIS element, λ is the wavelength, $d^i_{\mathbf{m},\mathbf{n}}$ is the distance between element $\mathbf{m}$ on TIS i and element $\mathbf{n}$ on TIS $i+1$, and $\phi^i_{\mathbf{m},\mathbf{n}}$ is the angle between the z-axis and the direction from element $\mathbf{m}$ on TIS i to element $\mathbf{n}$ on TIS $i+1$. Additionally, let $\gamma^i_{\mathbf{m}}$ represent the transmission coefficient of element $\mathbf{m}$ on TIS i.

Let $\mathbf{I}_i \in \mathbb{C}^{N_i\times N_i}$ and $\mathbf{O}_i \in \mathbb{C}^{N_i\times N_i}$ denote the input and output signals of TIS i, respectively. Moreover, let $I^i_{\mathbf{m}}$ and $O^i_{\mathbf{m}}$ represent the input and output signals of element $\mathbf{m}$ on TIS i, which are the $\mathbf{m}^{\text{th}}$ entries in $\mathbf{I}_i$ and $\mathbf{O}_i$, respectively. Given the input signal $\mathbf{I}_i$ of TIS i, the output signal $O^{i+1}_{[n_x,n_y]}$ of element $[n_x, n_y]$ on TIS $i + 1$ is given by

$$O^{i+1}_{[n_x,n_y]} = \sum_{m_x=-\frac{N_i-1}{2}}^{\frac{N_i-1}{2}} \sum_{m_y=-\frac{N_i-1}{2}}^{\frac{N_i-1}{2}} \Big(I^i_{[m_x,m_y]} \times \gamma^i_{[m_x,m_y]} h^i_{[m_x,m_y],[n_x,n_y]} \gamma^{i+1}_{[n_x,n_y]}\Big). \quad (2)$$

C. A General Overview of TIS-Based Convolutional Layer

As illustrated in Fig. 1, the three TISs function as a 2-D convolution operator, aiming to have the output of TIS 3, which is $\mathbf{O}_3$, emulate the convolutional results between the input signal $\mathbf{I}_1$ of TIS 1 and a specified $\mathbf{K}$ as follows

$$\mathbf{O}_3 \propto \mathbf{I}_1 * \mathbf{K}. \quad (3)$$

Information of the desired kernel $\mathbf{K}$ is mapped to TIS 2. Details of the system design will be introduced in Sec. III.

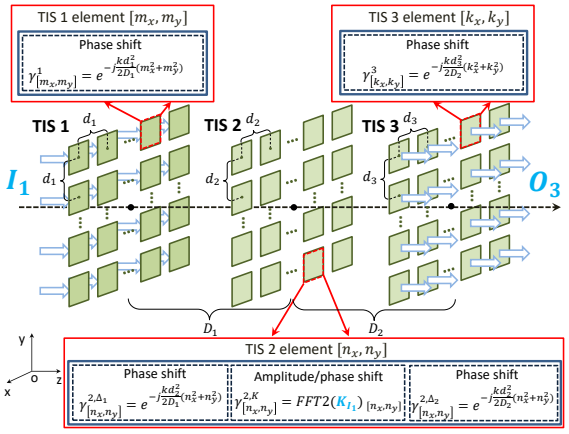


Fig. 1. Illustration of 2-D convolution operator with three TISs.

III. Design of 2-D Convolutional Layer Based on Transmissive Intelligent Surfaces

In what follows, we target a 2-D convolution operation of $\mathbf{I} * \mathbf{K} = \mathbf{O}$ where $\mathbf{I} \in \mathbb{C}^{N_I\times N_I}$, $\mathbf{K} \in \mathbb{C}^{N_K\times N_K}$, and $\mathbf{O} \in \mathbb{C}^{N_O\times N_O}$. Without loss of generality, it is assumed

that $N_I \geq N_K$. Moreover, a 2-D convolution operation without zero padding to $\mathbf{I}$ and $\mathbf{K}$ is considered, resulting in $N_O = N_I - N_K + 1$.

This section presents the design of the transmission coefficients for each element on the three TISs. These transmission coefficients depend on the distance between adjacent TISs and the spacing of TIS elements. In other words, while the size of the proposed structure can be adjusted, the 2-D convolution operation can still be realized, provided the transmission coefficients of the TIS elements are designed as specified in this section.

Our design of 2-D convolution in the RF domain is based on 2-D fast Fourier transform (FFT). The design uses the fact that the convolution of two signals corresponds to the element-wise product of their Fourier transforms, which will be discussed in Sec. III-A. Sec. III-B details the 2-D FFT operation using TISs, and Sec. III-C specifies the 2-D convolution in the RF domain.

A. 2-D Convolution Based on FFT

As mentioned above, the convolution operation of $\mathbf{I} * \mathbf{K} = \mathbf{O}$ can be achieved using the property of 2-D FFT as follows

$$\mathbf{O}' \propto FFT2(FFT2(\mathbf{I})FFT2(\mathbf{K}_I)), \tag{4}$$

where $\mathbf{O}'$ is the inverted version of $\mathbf{O}$ such that $\mathbf{O}'_{[x,y]} = \mathbf{O}_{[-x,-y]}$, $\mathbf{K}_I$ is the version of $\mathbf{K}$ with zero padding applied to match the dimensions of $\mathbf{I}$, and $FFT2(\cdot)$ represents 2-D FFT operation as follows

$$FFT2(\mathbf{X})_{[x,y]} = \sum_{u=\frac{1-N_x}{2}}^{\frac{N_x-1}{2}} \sum_{v=\frac{1-N_x}{2}}^{\frac{N_x-1}{2}} \mathbf{X}_{[u,v]} e^{-j\frac{2\pi}{N_x}(xu+yv)}. \tag{5}$$

where $\mathbf{X} \in \mathbb{C}^{N_x \times N_x}$. Before presenting the overall convolutional layer design, we present our design to implement a 2-D FFT using TISs.

B. 2-D FFT Based on TISs

Let $k = \frac{2\pi}{\lambda}$ denote the wavevector. Based on Fresnel approximation [12], the signal transmission model in (1) can be approximated as

$$h^i_{\mathbf{m},\mathbf{n}} \approx \frac{A_t e^{jkD_i}}{j\lambda D_i} e^{j\frac{k}{2D_i}\left((m_x d_i - n_x d_{i+1})^2 + (m_y d_i - n_y d_{i+1})^2\right)}, \tag{6}$$

when the following condition is satisfied

$$\sqrt{(m_x d_i - n_x d_{i+1})^2 + (m_y d_i - n_y d_{i+1})^2} \ll D_i. \tag{7}$$

Therefore, the output signal of element $\mathbf{n}$ on TIS $i+1$ given in (2) can be further approximated as follows

$$O^{i+1}_{[n_x,n_y]} = \frac{A_t e^{jkD_i}}{j\lambda D_i}\left(e^{j\frac{kd_{i+1}^2}{2D_i}(n_x^2+n_y^2)}\gamma^{i+1}_{[n_x,n_y]}\right) \times \sum_{m_x=-\frac{N_i-1}{2}}^{\frac{N_i-1}{2}} \sum_{m_y=-\frac{N_i-1}{2}}^{\frac{N_i-1}{2}} I^i_{[m_x,m_y]} e^{-j\frac{kd_i d_{i+1}}{D_i}(m_x n_x + m_y n_y)} \times \left(e^{j\frac{kd_i^2}{2D_i}(m_x^2+m_y^2)}\gamma^i_{[m_x,m_y]}\right). \tag{8}$$

By comparing (5) and (8), it can be observed that we have

$$O^{i+1}_{[n_x,n_y]} = \frac{A_t e^{jkD_i}}{j\lambda D_i} FFT2(\mathbf{I}_i)_{[n_x,n_y]}, \tag{9}$$

if condition (7) and the following conditions are satisfied

$$\gamma^{i+1}_{[n_x,n_y]} = e^{-j\frac{kd_{i+1}^2}{2D_i}(n_x^2+n_y^2)}, \tag{10}$$

$$\gamma^{i}_{[m_x,m_y]} = e^{-j\frac{kd_i^2}{2D_i}(m_x^2+m_y^2)}, \tag{11}$$

$$\frac{kd_i d_{i+1}}{D_i} = \frac{2\pi}{N_i}. \tag{12}$$

Therefore, we propose a 2-D FFT operator using two TISs with distance D_1, each comprising $N_1 \times N_1$ elements, as illustrated in Fig. 2. The transmission coefficient of element $[m_x, m_y]$ on TIS i ($i \in \{1,2\}$) is given by $\gamma^i_{[m_x,m_y]} = e^{-j\frac{kd_i^2}{2D_1}(m_x^2+m_y^2)}$. Moreover, the distance between the centers of the two TISs and the TIS element spacing should meet the following two conditions

$$D_1 = \frac{kd_1 d_2 N_1}{2\pi}, \tag{13}$$

$$\frac{\sqrt{2}}{2}(d_1+d_2)(N_1-1) \ll D_1, \tag{14}$$

where condition (14) comes from (7) to satisfy the condition of Fresnel approximation. Then, the output of TIS 2, which is $\mathbf{O}_2$, is determined by the 2-D FFT of the input of TIS 1, which is $\mathbf{I}_1$, as shown in (9).

C. 2-D Convolution Based on TISs

1) TIS transmission coefficient design: As shown in Sec. III-B, the output of TIS 2 forms a 2-D FFT plane of TIS 1's input by using the system design in Fig. 2. Therefore, if three TISs are placed sequentially as in Fig. 1, the inverted version of output $\mathbf{O}_3$ from TIS 3, denoted by $\mathbf{O}'_3$, is given by

$$\mathbf{O}'_3 \propto \frac{A_t e^{jkD_1}}{j\lambda D_1}\frac{A_t e^{jkD_2}}{j\lambda D_2} FFT2(FFT2(\mathbf{I}_1)FFT2(\mathbf{K}_{I_1})), \tag{15}$$

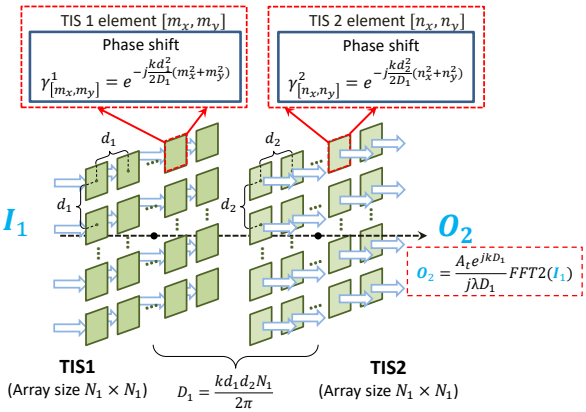


Fig. 2. Illustration of a 2-D FFT operator with two TISs.

where $\mathbf{K}_{I_1}$ is the version of kernel $\mathbf{K}$ with zero padding to match the dimension of $\mathbf{I}_1$, and the transmission coefficients of TIS 1 and TIS 3 are determined by

$$\gamma^1_{[m_x,m_y]} = e^{-\frac{jkd_1^2}{2D_1}(m_x^2+m_y^2)}, \tag{16}$$

$$\gamma^3_{[k_x,k_y]} = e^{-\frac{jkd_3^2}{2D_2}(k_x^2+k_y^2)}. \tag{17}$$

Furthermore, the transmission coefficient of TIS 2 is determined by three parts such that

$$\gamma^2_{[n_x,n_y]} = \gamma^{2,\Delta_1}_{[n_x,n_y]} + \gamma^{2,K}_{[n_x,n_y]} + \gamma^{2,\Delta_2}_{[n_x,n_y]}, \tag{18}$$

where

$$\gamma^{2,\Delta_j}_{[n_x,n_y]} = e^{-\frac{jkd_2^2}{2D_j}(n_x^2+n_y^2)}, j \in \{1,2\}, \tag{19}$$

$$\gamma^{2,K}_{[n_x,n_y]} = FFT2(\mathbf{K}_{I_1})_{[n_x,n_y]}. \tag{20}$$

Herein, $\gamma^1_{[m_x,m_y]}$ and $\gamma^{2,\Delta_1}_{[n_x,n_y]}$ are used to get the 2-D FFT of input $\mathbf{I}_1$, $\gamma^{2,K}_{[n_x,n_y]}$ is used to map $FFT2(\mathbf{K}_{I_1})$ to TIS 2, and $\gamma^{2,\Delta_2}_{[n_x,n_y]}$ and $\gamma^3_{[k_x,k_y]}$ are used to get the 2-D FFT of $FFT2(\mathbf{I}_1)FFT2(\mathbf{K}_{I_1})$. In other words, the output of TIS 3 is determined by the inverted version of $\mathbf{I}_1 * \mathbf{K}$, such that

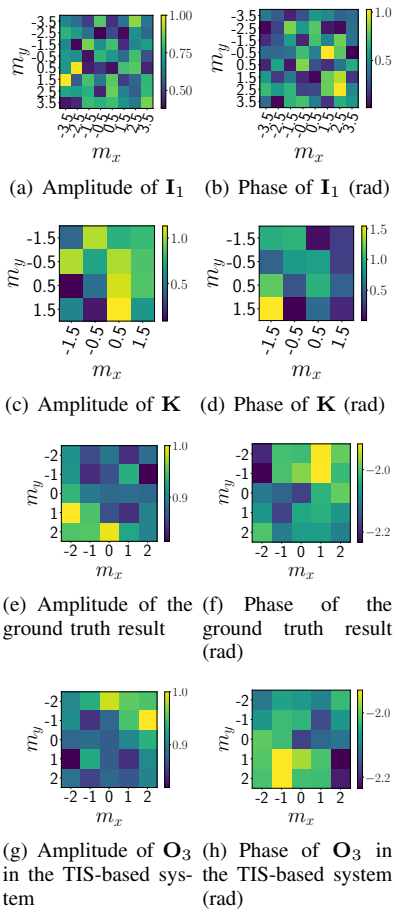
$$O^3_{[-k_x,-k_y]} \propto \frac{A_t e^{jkD_1}}{j\lambda D_1}\frac{A_t e^{jkD_2}}{j\lambda D_2}(\mathbf{I}_1 * \mathbf{K})_{[k_x,k_y]}. \tag{21}$$

To summarize, the design of the transmission coefficients for elements $[m_x, m_y]$ on TIS 1, TIS 2, and TIS 3 are provided in (16), (18), and (17), respectively. These coefficients are dependent on the distance between adjacent TISs and the spacing of the TIS elements. Furthermore, the choice of distance between adjacent TISs and the spacing of the TIS elements must satisfy conditions (7) and (12). In addition, the TIS size parameters are set as $N_1 = N_2 = N_I$ and $N_3 = N_O$.

2) Validation: To validate the TIS-based 2-D convolution, the three-TIS system illustrated in Fig. 1 based on the transmission model in (1) is simulated. Herein, the operating frequency is 60 GHz, D_1 and D_2 are both set as 1.5 m, and TIS element spacing parameters are $d_1 = d_3 = 0.02$ m and $d_2 = 0.046$ m. Moreover, the TIS array size parameters are set as $N_1 = N_2 = 8$ and $N_3 = 5$. Input signal $\mathbf{I}_1 \in \mathbb{C}^{8\times8}$ and $\mathbf{K} \in \mathbb{C}^{4\times4}$ with random amplitudes and phases are presented in Fig. 3(a), Fig.3(b), Fig.3(c) and Fig.3(d). The amplitude and phase of the ground truth result, $\mathbf{I}_1 * \mathbf{K}$, are shown in Fig. 3(e) and Fig. 3(f), and the amplitude and phase of the output signal $\mathbf{O}_3$ from TIS 3 are shown in Fig. 3(g) and Fig. 3(h). Note that the amplitudes in Fig. 3(e) and Fig. 3(g) are normalized so that the maximum amplitude is 1. It can be observed from Fig. 3 that the normalized output signal from TIS 3 is a good approximation of the inverted version of the normalized ground truth result.

D. Received Power Improvement

The Fresnel approximation condition, as discussed in Sec.III-B, may result in TISs being positioned at a distance which causes significant signal propagation loss. However, the



(a) Amplitude of $\mathbf{I}_1$ (b) Phase of $\mathbf{I}_1$ (rad)

(c) Amplitude of $\mathbf{K}$ (d) Phase of $\mathbf{K}$ (rad)

(e) Amplitude of the ground truth result (f) Phase of the ground truth result (rad)

(g) Amplitude of $\mathbf{O}_3$ in the TIS-based system (h) Phase of $\mathbf{O}_3$ in the TIS-based system (rad)

Fig. 3. Demonstration of input signal $\mathbf{I}_1$, kernel $\mathbf{K}$, and the convolutional results $\mathbf{O}_3$ from TIS-based system compared to the ground truth.

performance of OTA computation relies on a good signal-to-noise ratio (SNR) at the receiving antennas used to obtain the OTA results. Therefore, enhancing the power of the output $\mathbf{O}_3$ from TIS 3 is necessary. To improve the power of $\mathbf{O}_3$, we will increase the number of TIS elements of TIS 1 and TIS 2 without changing the convolutional results of $\mathbf{O}_3$. Let U_i be the TIS expansion rate for TIS i, which is defined as the ratio of the TIS dimension after expansion to its original dimension. The total number of elements of TIS i with expansion rate U_i is $(U_i N_i)^2$, and the element spacing is changed to $\hat{d}_i = \frac{d_i}{U_i}$. An illustration of $U_i = 3$ is shown in Fig. 4. It is noted that the maximum TIS expansion rate for TIS i is given by $U_i = \lfloor \frac{d_i}{d_e} \rfloor$ where $i \in \{1, 2\}$. The strategy of TIS element expansion for TIS 1 and TIS 2 is as follows:

- TIS 2: Since the plane where TIS 2 is located represents

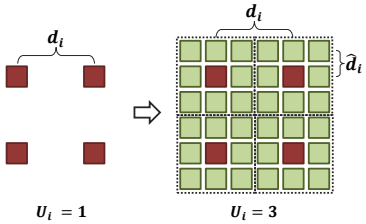


Fig. 4. Illustration of TIS expansion for the purpose of power improvement.

the Fourier transformation plane according to Sec.III-B, increasing the number of elements on TIS 2 results in increased sampling resolution of the Fourier spectrum, which will not impact the convolutional results. With TIS expansion U_2, the transmission coefficient $\gamma^2_{[n_x,n_y]}$ for TIS 2, where $n_x, n_y \in \{-\frac{U_2N_2-1}{2}, -\frac{U_2N_2-1}{2}+1, \cdots, \frac{U_2N_2-1}{2}-1, \frac{U_2N_2-1}{2}\}$, remains as given in (18), but with d_2 in (19) replaced by $\hat{d}_2$, and (20) refined as $\gamma^{2,K}_{[n_x,n_y]} = FFT2(\mathbf{K}_{I_1})_{[\frac{n_x}{U_2}, \frac{n_y}{U_2}]}$.

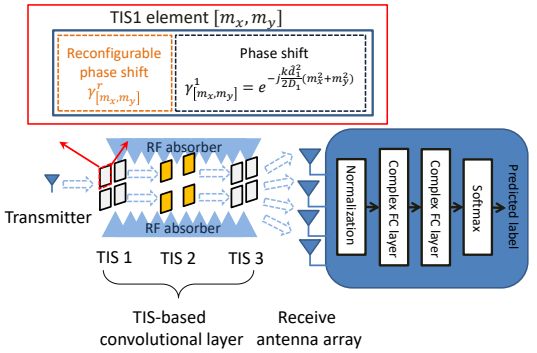
- TIS 1: Each $U_1 \times U_1$ element is grouped into a subarray, with all elements in a subarray receiving identical inputs as the input to one element before TIS expansion. This repetition of signals on TIS 1 distorts the 2-D FFT results generated by TIS 2. To offset this distortion, element $[n_x, n_y]$ on TIS 2 will have an additional amplitude attenuation, denoted as $\gamma^{2,A}_{[n_x,n_y]} = \min_{n_x,n_y}\{s(n_x,n_y)\}/s(n_x,n_y)$, where $s(n_x,n_y) = |FFT2(\mathbf{E}_{U_1})_{[\frac{n_x}{U_2N_2}, \frac{n_y}{U_2N_2}]}|$ and $\mathbf{E}_{U_1} \in \mathbb{R}^{U_1 \times U_1}$ is a matrix in which all elements are 1. Moreover, the transmission coefficient $\gamma^1_{[m_x,m_y]}$ for TIS 1, where $m_x, m_y \in \{-\frac{U_1N_1-1}{2}, -\frac{U_1N_1-1}{2}+1, \cdots, \frac{U_1N_1-1}{2}-1, \frac{U_1N_1-1}{2}\}$, remains as given in (16), but with d_1 replaced by $\hat{d}_1$.

IV. Simulation Results

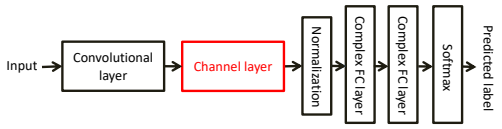
The TIS-based convolutional layer is implemented in simulation to evaluate the proposed system. We follow a common practice in the literature for validating new ML architectures by designing a neural network with our convolutional layer to perform an image classification task on handwritten digits.

A. Set-up

Without loss of generality, a simple TIS-based convolutional layer with one kernel implemented with three TISs is considered. In the preceding mathematical analyses, it was assumed that the channel between adjacent TISs is deterministic and known. One way of approximating this in practice would be to enclose the entire system in RF absorbing material, thereby eliminating all multipath components. This issue is discussed in more detail in Section V. To validate the proposed design, the UCI Digits dataset is used [13], which consists of 5620 samples of 8×8 images of written digits from $0-9$. To feed the input image to the TIS-based convolutional layer, the combination of a single-antenna transmitter and a programmable TIS is used. The size of the programmable TIS is $(8U_1) \times (8U_1)$, and a reconfigurable phase shift $\gamma^r_{[m_x,m_y]}$ is used to map the pixel value to the phase of transmission coefficients of elements on TIS 1, as illustrated in Fig. 5(a). If TIS expansion is used for TIS 1, all the reconfigurable phase shifts $\gamma^r_{[m_x,m_y]}$ within a $U_1 \times U_1$ subarray are identical. The transmitter only transmits a constant unmodulated signal, and the programmable TIS tunes the phase of $\gamma^r_{[m_x,m_y]}$ corresponding to the pixel value of the input image. The RF signals are then modified by the TIS-based convolutional layer, before being received by a 4-antenna array at the receiver. The



(a) Neural network with a TIS-based convolutional layer



(b) Neural network with a classic convolutional layer

Fig. 5. Structure of neural networks with standard convolutional layer and TIS-based convolutional layer.

received signal is further processed with two complex-valued fully-connected (FC) layers and a softmax layer implemented inside the receiver's processor. The full pipeline is shown in Fig. 5(a).

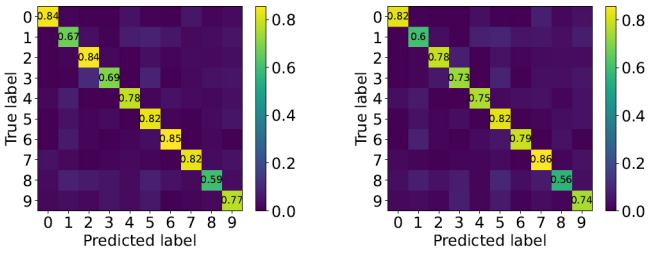
A neural network with a classic complex-valued convolutional layer with a 4×4 kernel based on [14], as illustrated in Fig. 5(b), is trained in PyTorch. The untrainable channel layer in Fig. 5(b) is used to model signal propagation from TIS 3 to the receive antenna array. After convergence is reached in training, the trained weights of the kernel are extracted and then mapped to the transmission coefficients of TIS 2 as detailed in Sec. III.

In what follows, the operating frequency is 60 GHz. The TIS array size parameters are set as $N_1 = N_2 = 8$ and $N_3 = 5$. The distance between TIS 1 and TIS 2, and the distance between TIS 2 and TIS 3 are assumed to be the same, which is denoted by D. The element spacing of TIS 1 and TIS 3 are $d_1 = d_3 = 0.02$ m. The distances between the transmitter and the center of TIS 1, and between the center of the receive antenna array and the center of TIS 3, are both set at 0.1 m. Additionally, the size of every TIS element is $A_t = \frac{\lambda}{2} \times \frac{\lambda}{2}$.

B. Validating the TIS-Based Convolutional Layer

To show that the proposed TIS-based convolutional layer has similar behavior as the classic convolutional layer, the performance of the digit classification task is compared between the TIS-based and the classic designs. Herein, the distance between adjacent TISs is $D = 0.7$ m, with element spacing of TIS 2 as $d_2 = 0.0218$ m according to (12). The TIS expansion rates for TIS 1 and TIS 2 are chosen as $U_1 = U_2 = 7$. Moreover, the transmit power is set as 15 dBm.

The confusion matrix of the classification results on the test data is demonstrated in Fig. 6. According to Fig. 6, there is only a slight gap between the classification accuracy of the proposed TIS-based design and the classic design, which



(a) Neural network with a classic convolutional layer

(b) Neural network with a TIS-based convolutional layer

Fig. 6. Confusion matrices.

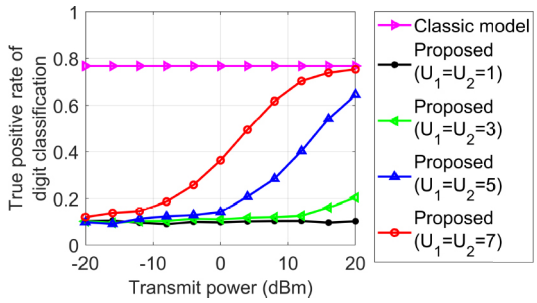


Fig. 7. True positive rate vs. transmit power for various TIS expansion rates ($D = 0.7$ m, $d_1 = 0.02$ m).

demonstrates the efficacy of the proposed TIS-based solution as a good substitute for the classic convolutional layer. This slight gap between the proposed and classic designs primarily arises from two aspects:

1) The proposed TIS-based solution is based on the Fresnel approximation, which requires adherence to the condition of distance between consecutive TISs as discussed in Sec. III-B. Considering the size of the proposed structure, a reasonable selection of the distance between TISs is used in this simulation, resulting in a slight variance between the accurate channel and its Fresnel approximation. If an increased distance between TISs is allowed, along with a higher transmit power to compensate for the larger path loss caused by this distance, a more precise Fresnel approximation can be obtained to improve the classification performance.
2) The training process of classic neural networks has no consideration of noise at the receive antennas, leading to an ideal classification performance. Nonetheless, in the proposed TIS-based system, classification performance relies on SNR at the receive antennas. Therefore, the noise has an impact on performance in the proposed TIS-based system.

C. Classification Accuracy vs. Transmit Power with Adjusted TIS Expansion Rate

An increase in transmit power can enhance the accuracy of TIS-based convolutional operations, since the SNR at the receive antennas is improved. The strategy proposed in Sec. III-D can improve SNR without increasing transmit power by increasing the number of TIS elements. In this section, the relationship between transmit power and the true positive rate of digit classification task is evaluated using the received power improvement strategy proposed in Sec. III-D.

According to Fig. 7, when the TIS expansion rate is $U_1 = U_2 = 1$ for the TIS-based system, the true positive rate remains relatively unchanged as the transmit power increases from -20 dBm to 20 dBm. This observation is attributed to the limited number of TIS elements, which constrains the received power at the antennas. As U_1 and U_2 increase from 1 to 7, there is a significant improvement of true positive rate, which demonstrates the effectiveness of the proposed received power improvement strategy. When $U_1 = U_2 = 7$, the true positive rate almost reaches the maximum value provided by the classic model as the transmit power increases to above 16 dbm. The results demonstrate a trade-off between transmit power and hardware complexity. To maintain classification accuracy, the number of TIS elements can be increased to reduce transmit power.

D. Classification Accuracy vs. Transmit Power with Adjusted Distance Between TISs

The distance between two adjacent TISs, which is D, is an important metric in TIS-based convolutional layer in that:

1) this distance directly influences the sizes of the proposed design, thereby affecting the feasibility of real-world implementation, and
2) this distance is related to the Fresnel approximation, as discussed in Sec. III-B and, therefore, a wider range of feasible distances between TISs enables a more accurate approximation between the convolutional results using the proposed design and the ground truth.

As D changes, the element spacing d_2 for TIS 2 needs adjustment according to (12), which impacts the maximum TIS expansion rate supported by TIS 2. Herein, it is assumed that d_2 is determined by (12), and the maximum TIS expansion rate for TIS 2 is utilized as D varies. Fig. 8 shows the true positive rate for different distances with $U_1 = 7$. According to Fig. 8, there is hardly any reduction in the true positive rate as D decreases from 0.7 m to 0.5 m. For $D = 0.3$ m, there is a slight decrease in the true positive rate but the proposed design still achieves approximately 92% of the ideal true positive rate of the classic model in this case. However, a reduction to $D = 0.2$ m leads to a significant performance reduction when the proposed design is used. This result demonstrates that slight violations of the Fresnel approximation can still yield satisfying classification performance, while allowing for a reduction in the size of the proposed design.

V. Discussion

In this section, we discuss the takeaways, limitations, and further improvements that can be made in future research.

The key finding in this work is that a 2-D convolutional layer can be constructed in the RF domain using cascaded TISs. To the authors' best knowledge, this is the first work that enables an OTA 2-D convolutional layer. And compared to the previous attempt to create an OTA convolutional layer [6],

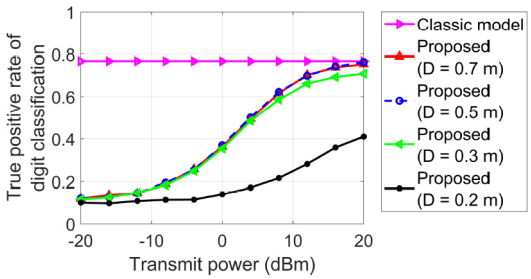


Fig. 8. True positive rate vs. transmit power for various distances between TISs ($U_1 = 7$, $d_1 = 0.02$m).

which uses multiple transmitters and multiple IRSs to create a reconfigurable multi-tap 1-D convolution directly in the time domain, our proposed method removes the need for tight synchronization. In addition, the size of input data and the convolution kernel size can be easily adjusted by simply using different numbers of TIS elements, instead of needing to increase the number of TISs. This could potentially allow for better flexibility and scalability in practical deployments.

This work has a few limitations: (1) it is assumed that the whole RF pipeline is enclosed in RF absorbing materials, which eliminate the undesirable multipath components, and that the channel information is perfectly known. Designing a 2-D convolutional layer when the measurement of the channel is imperfect or unavailable remains an open problem. To address similar problems, other works have adopted discrete optimization or reinforcement learning approaches [7] to train the weights in an online fashion. Such approaches are beyond the scope of this initial work and are an open problem for future research. (2) The TISs incur added cost and complexity due to the additional hardware. Recent progress in low-complexity designs with 1-bit and 2-bit phase shifts, e.g. in [15], [16], points the way for additional open topics in this area. Furthermore, if reconfiguration of the ML model is not required, non-reconfigurable TIS elements can be employed to reduce hardware complexity, as this eliminates the need for controlling hardware to adjust TIS transmission coefficients, which is typically required in most existing studies on intelligent surfaces. (3) Power consumption of a TIS-based convolutional layer primarily arises from the transmitter on the edge device, as the TIS itself offers a low-power advantage due to its composition of passive electromagnetic elements. To achieve better performance, there is a trade-off between transmit power and hardware complexity. As discussed in Sec. IV-C, an increase in the number of TIS elements is needed to reduce transmit power while maintaining classification accuracy. Therefore, a careful consideration of both power consumption and hardware complexity will be critical for real-world applications.

VI. Conclusion

In this paper, we introduced an RF-based convolutional layer using three TISs, which represents the first structure capable of implementing 2-D convolutional operations using OTA computation. To validate the proposed design, a simple neural network with a single convolutional layer and one kernel was trained for a digit classification task. Simulation results showed that, given properly designed TIS parameters, the neural network with the proposed TIS-based convolutional layer achieves a classification accuracy very close to the classic complex-valued convolutional layer. However, to achieve good performance, the number of TIS elements and the distance between TISs must be properly selected, which raises issues of trade-offs between power consumption and hardware complexity that require further investigation to demonstrate use in practical real-world applications.

References

[1] R. Singh and S. S. Gill, "Edge AI: A survey," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 71–92, 2023.

[2] P. P. Ray, "A review on TinyML: State-of-the-art and prospects," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, pp. 1595–1623, 2022.

[3] X. Ma, Y. Zhao, L. Zhang, Q. Gao, M. Pan, and J. Wang, "Practical device-free gesture recognition using wifi signals based on metalearning," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 228–237, 2020.

[4] C. Xiao, Y. Lei, Y. Ma, F. Zhou, and Z. Qin, "Deepseg: Deep-learning-based activity segmentation framework for activity recognition using wifi," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5669–5681, 2021.

[5] G. Reus-Muns, K. Alemdar, S. G. Sanchez, D. Roy, and K. R. Chowdhury, "AirFC: Designing fully connected layers for neural networks with wireless signals," in *Proceedings of the Twenty-Fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2023, pp. 71–80.

[6] S. G. Sanchez, G. Reus-Muns, C. Bocanegra, Y. Li, U. Muncuk, Y. Naderi, Y. Wang, S. Ioannidis, and K. R. Chowdhury, "AirNN: Over-the-air computation for neural networks via reconfigurable intelligent surfaces," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 2470–2482, 2022.

[7] C. Liu, Q. Ma, Z. J. Luo, Q. R. Hong, Q. Xiao, H. C. Zhang, L. Miao, W. M. Yu, Q. Cheng, L. Li *et al.*, "A programmable diffractive deep neural network based on a digital-coding metasurface array," *Nature Electronics*, vol. 5, no. 2, pp. 113–122, 2022.

[8] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Communications Surveys Tutorials*, vol. 23, no. 3, pp. 1546–1577, 2021.

[9] X. Yuan, Y.-J. A. Zhang, Y. Shi, W. Yan, and H. Liu, "Reconfigurable-intelligent-surface empowered wireless communications: Challenges and opportunities," *IEEE Wireless Communications*, vol. 28, no. 2, pp. 136–143, 2021.

[10] Y. Liu, X. Mu, J. Xu, R. Schober, Y. Hao, H. V. Poor, and L. Hanzo, "STAR: Simultaneous transmission and reflection for 360° coverage by intelligent surfaces," *IEEE Wireless Communications*, vol. 28, pp. 102–109, 2021.

[11] J. An, M. Di Renzo, M. Debbah, and C. Yuen, "Stacked intelligent metasurfaces for multiuser beamforming in the wave domain," in *Proceedings of the IEEE International Conference on Communications*, 2023, pp. 2834–2839.

[12] D. Mas, J. Garcia, C. Ferreira, L. M. Bernardo, and F. Marinho, "Fast algorithms for free-space diffraction patterns calculation," *Optics Communications*, vol. 164, pp. 233–245, 1999.

[13] E. Alpaydin and C. Kaynak, "Optical recognition of handwritten digits," UCI Machine Learning Repository, 1998, DOI: https://doi.org/10.24432/C50P49.

[14] M. W. Matthès, Y. Bromberg, J. de Rosny, and S. M. Popoff, "Learning and avoiding disorder in multimode fibers," *Physical Review X*, vol. 11, pp. 1–12, 2021.

[15] J. Zhang, R. Xiong, J. Liu, T. Mi, and R. C. Qiu, "Design and prototyping of transmissive ris-aided wireless communication," 2024. [Online]. Available: http://arxiv.org/abs/2402.05570

[16] J. Tang, S. Xu, and F. Yang, "Design of a 2.5-d 2-bit reconfigurable transmitarray element for 5g mmwave applications," in *2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting*, 2020, pp. 631–632.