

**ORIGINAL ARTICLE**

International Journal of Applied Mathematics in Control Engineering

Journal homepage: <http://www.ijamce.com>

# Multi-Modal Physically Constrained Network for 3D Reconstruction of Complex Occluded Scenes in Terahertz Imaging and Its Applications

Qi Cheng<sup>1\*</sup> | Tichen Wang<sup>1</sup> | Ziyin Liu<sup>1</sup> | Yibo Cheng<sup>1</sup> | Qiulin Chen<sup>1</sup>

<sup>1</sup>Mechanical and Automotive Engineering,  
Qingdao University of Technology, Qinnngdao,  
Shandong, 266520, China

**Correspondence**

Qi Cheng PhD, Mechanical and Automotive  
Engineering, Qingdao University of Technology,  
Qinnngdao, Shandong, 266520, China  
Email: [chengqi@qut.edu.cn](mailto:chengqi@qut.edu.cn)

**Article Info**

Article history:

Received 4 September 2025

Accepted 19 November 2025

Available online 21 November 2025

**Abstract**

Terahertz (THz) holographic imaging is hindered by occlusion and the lack of physical consistency in data-driven reconstruction. To address this, we propose a physics-constrained deep learning framework that explicitly embeds THz wave propagation laws—based on angular spectrum theory and Fresnel diffraction—into network design and training. We construct a dynamic multi-modal dataset simulating realistic occlusion across biomedical, industrial, and security scenarios, and introduce MT-JointNet, a multi-task network that jointly reconstructs object sequences and enforces physical plausibility via a diffraction-consistency loss and a physics-guided attention mechanism. A self-training strategy with physics-based pseudo-label filtering further enhances generalization. Experiments show high-fidelity reconstructions with strong adherence to physical laws (consistency score  $> 0.88$ ), even under severe multi-object occlusion, demonstrating the effectiveness of tightly integrating physics into deep THz imaging.

**KEYWORDS**

Terahertz imaging, Occluded object reconstruction, Physics-driven deep learning, Multi-modal fusion, Multi-view self-training

## 1 | INTRODUCTION

Terahertz (THz) radiation, spanning the frequency range of 0.1 THz to 10 THz with corresponding wavelengths between 3 mm and 30  $\mu\text{m}$ , occupies a unique position in the electromagnetic spectrum. Terahertz holographic imaging technology has emerged as a promising non-destructive detection method due to its distinctive advantages: strong penetration through non-metallic and non-polar materials (e.g., ceramics, plastics, and fabrics), low photon energy avoiding ionizing damage to biological tissues, and unique molecular "fingerprint" recognition capability for substance identification[1, 2, 3]. These characteristics have driven its wide application prospects in industrial non-destructive testing[4], biomedical imaging[5], and security screening fields[6].

However, occlusion remains a critical bottleneck for practical terahertz imaging[7, 8]. In industrial, biomedical, and security

applications, targets are often hidden behind casings, tissues, or packaging layers. Traditional back-propagation-based methods[9] suffer from mutual interference among occluded objects and depend heavily on hard-to-acquire phase information, resulting in blurred edges and structural distortions[10]. Although physics-driven deep learning has improved reconstruction accuracy, significant challenges persist in complex real-world scenarios.

In recent years, research on terahertz occluded object reconstruction has evolved along two main paths: physics-model-based methods[11, 12] and deep learning-based methods[13]. Traditional physics-model-based approaches, represented by angular spectrum theory[14] and Fresnel diffraction formula[15], simulate the propagation process of terahertz waves to invert object structures. However, these methods suffer from high computational complexity and are sensitive to noise, making them ineffective for multi-object complex occlusion scenarios. With the development of deep learning, convolutional neural networks (CNNs) have been widely applied in terahertz imaging due to their powerful feature extraction capabilities. A previous study proposed a physics-driven self-training neural network, which uses angular spectrum theory to generate synthetic datasets for supervised pre-training and introduces a small amount of unlabeled experimental data for self-training, effectively reducing mutual interference between two occluded objects[16]. However, this method has obvious limitations: the synthetic data only based on MNIST datasets and simple planar metal objects has poor generalization to real complex scenes; the physical constraints are only introduced through data generation rather than dynamic integration into network training, leading to non-physical reconstruction results; and the reliance on single-modal amplitude data and single-view imaging makes it difficult to handle multi-object overlapping occlusion[17].

In related fields, multi-modal and multi-view fusion technologies have shown great potential for improving occlusion reconstruction performance. In terahertz MIMO imaging, multi-antenna technology is used to collect multi-view signals to suppress multi-path effects and improve imaging resolution[18]. In single-pixel phase imaging, embedding physical layers (Fresnel diffraction layer and SPI layer) into neural networks and designing composite loss functions have achieved high-quality reconstruction of phase objects[19]. In terahertz tomography, multi-view subspace attention networks that fuse spatio-spectral features have significantly improved the restoration quality of occluded objects[20]. However, the integration of these advanced technologies into terahertz holographic reconstruction of complex occluded scenes remains insufficiently explored.

This study aims to address three critical limitations in current data-driven terahertz holographic imaging: (1) poor generalization under complex multi-object occlusion due to oversimplified synthetic data, (2) lack of dynamic physical constraints leading to non-physical reconstructions, and (3) limited applicability to realistic dynamic scenarios. To this end, we first construct a dynamic, multi-modal, and multi-scene THz dataset that integrates biomedical, industrial, and security inspection contexts—incorporating amplitude, polarization, and near-infrared modalities along with realistic motion trajectories—thereby moving beyond the static, single-scene, MNIST-like paradigms of prior work. Second, we propose MT-JointNet, a physics-constrained multi-task network that explicitly embeds angular spectrum theory and Fresnel diffraction into its architecture. This framework jointly reconstructs dynamic object sequences and predicts spatial correlation parameters, while enforcing physical plausibility through a diffraction-consistency loss and a physics-guided attention mechanism. Third, we advance the self-training paradigm via a multi-view fusion strategy that leverages temporal coherence and physics-based pseudo-label filtering to enhance robustness without requiring extensive labeled experimental data. Our approach demonstrates high-fidelity reconstruction, strong adherence to physical laws, and effective handling of severe occlusion in synthetic yet physically realistic THz sequences, laying a foundation for future deployment in dynamic industrial inspection, biomedical diagnosis, and security screening.

## 2 | METHODOLOGY

### 2.1 | Multi-Scene Multi-Modal Dataset Construction

#### *Dataset Composition and Acquisition*

As shown in Fig. 1, this study builds a terahertz dynamic imaging time-series multimodal dataset based on MNIST. It randomly selects 2 digits, generates motion/rotation trajectories to simulate dynamics, and creates 5-frame sequences with superimposed digits and dynamic occlusion masks. The 5-channel (amplitude, 3 polarizations, near-infrared) terahertz data,



the key data foundation for model training. Among these, the phase information is not directly used as an input channel; instead, it serves as a supervisory signal to constrain the output of the model's diffraction module, thereby ensuring the rationality and accuracy of the physical simulation. The amplitude, on the other hand, is directly fed into the model as an independent channel within the multimodal data, constituting a core feature dimension that carries the essential physical information reflected by the superposition of the terahertz beam and the digital target. The amplitude (1 channel), polarization (3 channels), and near-infrared (1 channel) are concatenated in a fixed order to form a five-channel tensor sequence, which serves as the core primary input for the multi-temporal joint network MT-JointNet. Additionally, auxiliary inputs such as the spatial relationship matrix are fed into the network in parallel with this five-channel tensor, rather than being embedded within it. This maintains a modular design and supports independent modeling of multimodal features.

Each time series multimodal data and spatial location relationship matrix data are then fed as input into two MT-JointNet as shown in Fig. 1(e). MT-JointNet is a deep learning network designed for dynamic terahertz multimodal sequence processing. The network extracts multimodal features and temporal features separately through 3D convolution modules, accomplishes the joint fusion of multi-dimensional features via an attention mechanism, and then synchronously outputs the sequence reconstruction results and spatial correlation matrix of the target (see Appendix A for a detailed description). Meanwhile, it integrates a physical constraint module based on the diffraction law of terahertz waves and a self-training enhancement mechanism. It can not only address problems such as edge blurring and insufficient resolution in traditional terahertz imaging caused by long wavelengths but also reduce the interference of noise on imaging quality through inter-frame temporal consistency constraints. The outputs of the MT-JointNet are the reconstructed dynamic terahertz image sequence and spatial correlation matrix. In terahertz sequence processing, Dynamic MT Loss is specifically designed to quantify the discrepancy between MT-JointNet's dual outputs (sequence reconstruction results and spatial correlation matrix) and the ground truth.

### Scenario-based Physical Constraint Mechanism

The generation of dynamic terahertz multimodal data, the foundation of MT-JointNet's training and application, is rooted in the angular spectrum theory and Fresnel diffraction integral, which are extended to time-varying scenes involving object motion, rotation, and occlusion—key characteristics common in security checkpoints (moving luggage), industrial production lines (conveyor-belt parts), and medical diagnosis (dynamic skin monitoring). For a monochromatic terahertz wave with wavelength  $\lambda$ , the diffraction field  $U(x, y, z)$  at propagation distance from the object plane is given by the angular spectrum method, shown in Eq(1)[19]:

$$U(x, y, z) = \mathcal{F}^{-1} \left\{ \mathcal{F} \{U_0(x, y, 0)\} \cdot \exp \left[ jkz \sqrt{1 - (\lambda f_x)^2 - (\lambda f_y)^2} \right] \right\} \quad (1)$$

where  $\mathcal{F}$  and  $\mathcal{F}^{-1}$  denote the Fourier transform and inverse Fourier transform, respectively;  $f_x, f_y$  are spatial frequencies;  $k = \frac{2\pi}{\lambda}$  is the wave number. This formula accurately describes the propagation law of terahertz waves, laying a physical foundation for the authenticity of data in scenarios such as detecting hidden contraband in moving luggage or capturing dynamic defect changes in industrial parts.

To reduce the computational overhead in dynamic scenes and achieve the simulation of high-resolution sequences, paraxial approximation ( $z \gg \lambda$  and small angles) was carried out. The phase term simplifies to  $\exp(jkz - j\pi\lambda z(f_x^2 + f_y^2))$  and reducing the diffraction integral to the Fresnel approximation[21]:

$$U(x, y, z) = \frac{e^{jkz}}{j\lambda z} \iint U_0(x', y', 0) \exp \left( \frac{jk}{2z} [(x - x')^2 + (y - y')^2] \right) dx' dy' \quad (2)$$

where  $U_0(x', y', 0)$  denotes the complex amplitude at point  $(x', y')$  on the object plane;  $\exp \left( \frac{jk}{2z} [(x - x')^2 + (y - y')^2] \right)$  describes the phase change of the wave emitted from this point propagating to the point  $(x, y, z)$ . This simplified model balances computational efficiency and physical accuracy, making it suitable for real-time processing requirements in practical scenarios.

In dynamic scenes, the object plane  $U_0(x, y, t)$  evolves with time  $t$  under the combined effects of three interconnected factors: object motion, rotation, and dynamic occlusion. Specifically, object motion is characterized by displacements  $(\Delta x(t), \Delta y(t))$ ,

which are governed by the following kinematic equations:

$$\begin{cases} \Delta x(t) = v_x t + \frac{1}{2} a_x t^2 \\ \Delta y(t) = v_y t + \frac{1}{2} a_y t^2 \end{cases} \quad (3)$$

where  $\mathbf{v} = (v_x, v_y)$  denotes the velocity vector of the object, and  $\mathbf{a} = (a_x, a_y)$  represents its acceleration vector. For the rotational behavior, the rotation angle  $\theta(t)$  increases linearly from  $0^\circ$  to a maximum value  $\theta_{\max}$  over the total sequence length  $T$ , which is mathematically expressed as:

$$\theta(t) = \theta_{\max} \cdot \frac{t}{T} \quad (4)$$

To accurately depict the complex amplitude distribution of the initial wavefront under the synergy of time-varying objects and dynamic occlusions, the object plane is formulated as:

$$U_0(x', y', t) = U_{\text{obj}}(x - \Delta x(t), y - \Delta y(t), \theta(t)) \cdot (1 - O(x, y, t)) + U_{\text{occl}}(x, y, t) \quad (5)$$

In Eq. (5),  $O(x, y, t)$  is the time-varying occlusion mask defined by transmittance (ranging from 0 to 1, where 1 indicates full occlusion and 0 denotes no occlusion),  $U_{\text{obj}}$  denotes the complex amplitude field of the target object, and  $U_{\text{occl}}$  represents the complex amplitude field of the occluding medium—e.g., clothes covering contraband, packaging materials wrapping industrial parts, or adjacent skin tissue overlapping a lesion. Ultimately, 20,000 dynamic terahertz multimodal data are generated by systematically varying the motion parameters (i.e., velocity, acceleration, and maximum rotation angle), ensuring comprehensive coverage of diverse dynamic scenarios for model training and validation.

The object-plane complex amplitude  $U(x, y, t)$  in dynamic scenes evolves with time under the coupled effects of object motion, rotation, and dynamic occlusion, as formulated in the preceding equations. On this basis, terahertz multi-modal data—encompassing amplitude, polarization, and near-infrared (NIR) modalities—was further simulated by fusing the derived diffraction field information and inherent scene characteristics. The specific implementation methodology for each modal channel is detailed as follows: The amplitude modal data  $A(x, y, t)$  is directly derived from the modulus of the terahertz diffraction field  $U(x, y, t)$  (evolved from  $U_0(x, y, t)$  at spatial coordinates  $(x, y)$  and time  $t$ , i.e.,  $A(x, y, t) = |U(x, y, t)|$ ). The polarization modal data consists of three mutually orthogonal components. For the  $i$ -th component ( $i=1, 2, 3$ ), corresponding to three orthogonal polarization directions), the signal  $A_p^i(x, y, t)$  is obtained by weighting the amplitude data  $A(x, y, t)$  with the respective polarization efficiency  $\eta_i$ , expressed as [20]:

$$A_p^i(x, y, t) = A(x, y, t) \cdot \eta_i \quad (6)$$

The near-infrared modal data  $N(x, y, t)$  was simulated to incorporate both the shielding effect of occluders (consistent with the  $O(x, y, t)$  definition in dynamic scenes) and practical measurement noise, calculated by:

$$N(x, y, t) = A(x, y, t) \cdot (1 - 0.2 O(x, y, t)) + \varepsilon(t). \quad (7)$$

Here,  $O(x, y, t)$  denotes the dynamic occlusion mask at time  $t$  (value range:  $[0, 1]$ ; 0 = no occlusion, 1 = complete occlusion), which maintains consistency with the occlusion characterization in the object-plane modeling. The coefficient 0.2 quantifies the weak attenuation effect of occluders on NIR signals (e.g., thin clothing in security checks or translucent skin layers in medical imaging), and  $\varepsilon(t)$  represents additive Gaussian noise with a standard deviation of  $\sigma = 0.01$ , introduced to mimic random interference in actual terahertz detection systems.

By repeating this cycle for many pairs of objects, a portfolio of labeled sequential multimodality diffraction patterns for object pairs is generated. It is worth noting that we chose a fixed distance of 20 cm between the object and the detector for all pairs of the training dataset.

We have developed a self-training module integrated into the training pipeline of a single MT-JointNet model. Its core idea is to incorporate unlabeled, simulated THz dynamic sequence data into training, even when corresponding ground truth (GT) is unavailable or imprecise—such as inaccurate object localization (longitudinal/lateral) and ill-defined target geometric shapes. To

tackle these issues, the module uses the trained MT-JointNet to generate pseudo-labels for unlabeled simulated data, filtering high-confidence samples via the physics consistency score. Selected pseudo-labeled samples are mixed with labeled training data at a controlled ratio, boosting the model's generalization without high-precision manual annotations.

In the first step of self-training for a single MT-JointNet model (a CNN-based network), it is trained only on labeled simulated THz dynamic sequence data. The model is then applied to unlabeled simulated THz data to generate predictions, which are used as pseudo-labels after filtering high-confidence samples via a physics consistency score. These pseudo-labeled samples are mixed with a controlled ratio of the original labeled simulated data (instead of new labeled data subsets) for iterative retraining of the MT-JointNet. While semi-supervised learning typically targets scenarios with large volumes of unlabeled data (e.g., internet-sourced data), our THz imaging scenario involves an inverse setup: we use far more synthetic labeled THz dynamic sequence data (generated based on MNIST) than unlabeled simulated data. This design ensures the labeled synthetic data act as "guardrails" for stable training convergence. Additionally, mixing pseudo-labeled data at a controlled ratio (1/4 of batch size) instead of using the full unlabeled dataset accelerates training efficiency.

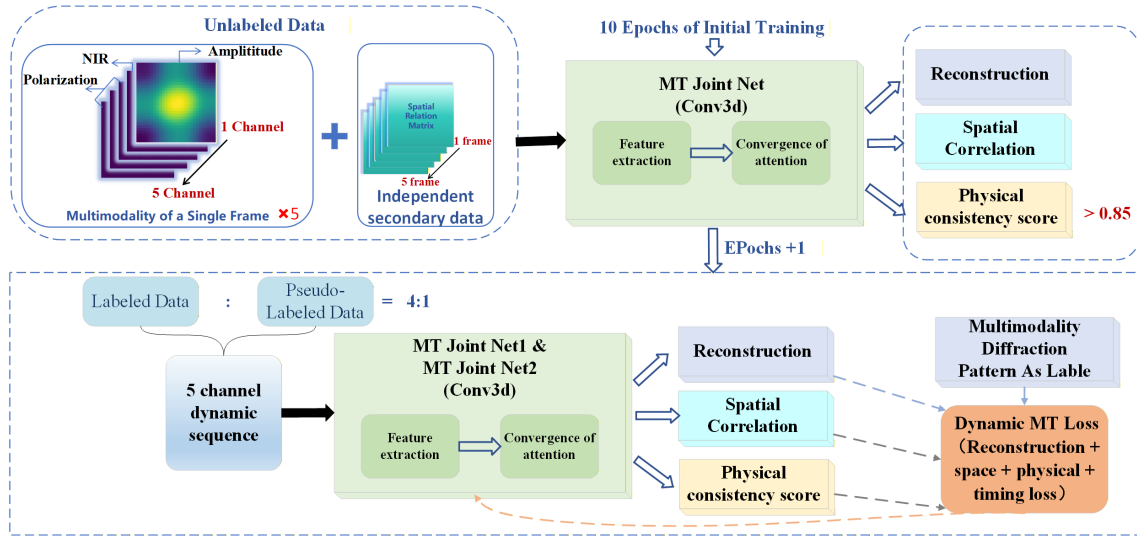


**FIGURE 2** Schematic process of the proposed object reconstruction algorithm indicates that there are additional self-learning steps for the images obtained from terahertz experiments.

As illustrated in Fig. 2, the self-training procedure is implemented as follows. The network first undergoes 10 epochs of initial training with labeled synthetic THz sequences generated from dynamic MNIST scenes. Afterward, the network parameters are used as initial values for self-training. The input remains MNIST-derived multi-modal THz sequence data, including: 1) Labeled data: dynamic THz sequences (distinct batches from the initial training data); 2) Unlabeled data: additional MNIST-derived dynamic THz sequences (no experimental data or Gaussian noise augmentation). High-confidence unlabeled samples are filtered via a physics constraint network (threshold = 0.85) to generate pseudo-labels, which are mixed with labeled data at a 1:4 ratio for joint training.

To ensure physical plausibility, we introduce a physics consistency score (PCS) that quantifies how well the reconstructed object adheres to THz diffraction laws. Specifically, PCS is defined as the normalized cross-correlation between the simulated hologram  $U_{\text{obs}}(x, y)$  and the field obtained by forward-propagating the reconstructed object plane  $\hat{U}_0(x', y')$  using the angular spectrum method:

$$\text{PCS} = \frac{\langle U_{\text{obs}}, \mathcal{P}[\hat{U}_0] \rangle}{\|U_{\text{obs}}\|_2 \cdot \|\mathcal{P}[\hat{U}_0]\|_2} \quad (8)$$

where  $\mathcal{P}[\cdot]$  denotes the forward diffraction operator, and  $\langle \cdot, \cdot \rangle$  is the complex inner product. During self-training, only samples with  $\text{PCS} \geq 0.85$  are retained as high-confidence pseudo-labels.

Unlabeled 5-frame 5-channel multimodal THz sequences are fed into the MT-JointNet that has completed 10 epochs of

training, which outputs reconstructed sequences. We calculate a physics consistency score by comparing the diffraction patterns of the reconstructed sequences and the dataset-native ground-truth (GT) sequences corresponding to the unlabeled data. If the score exceeds 0.85, the unlabeled data are assigned their native GT as pseudo-labels, thus becoming pseudo-labeled data for subsequent joint training. To ensure the model's outputs meet the diverse requirements of different application scenarios—such as high pixel-level reconstruction accuracy for biomedical diagnosis, reliable spatial relationship modeling for security inspection object classification, and strong temporal consistency for high-speed industrial detection—MT-JointNet integrates three tailored loss terms:  $L_1$  loss for pixel-level reconstruction error  $L_{rec}$ , MSE loss for spatial correlation matrix deviation  $L_{spatial}$ , and a temporal consistency penalty term  $L_{temp}$ . The total loss function, which balances these complementary constraints, is formulated as:

$$L_{total} = \alpha \cdot L_{rec} + \beta \cdot L_{spatial} + \gamma \cdot L_{temp} \quad (9)$$

where  $\alpha = 0.6$  (prioritizing reconstruction accuracy, a key demand for terahertz imaging in medical and industrial scenarios),  $\beta = 0.25$  (balancing spatial relationship modeling to support object recognition in security checks), and  $\gamma = 0.15$  (supplementing temporal consistency constraints for dynamic industrial and medical monitoring) are weights determined by cross-validation. The precise mathematical expressions for  $L_{rec}$ ,  $L_{spatial}$ , and  $L_{temp}$  are given below:

$$L_{rec} = \frac{1}{B \times T \times H \times W} \sum_{b=1}^B \sum_{t=1}^T \sum_{h=1}^H \sum_{w=1}^W \left| \hat{I}_{b,t,1,h,w} - I_{b,t,1,h,w}^{gt} \right| \quad (10)$$

$$L_{spatial} = \frac{1}{B \times N \times N} \sum_{b=1}^B \sum_{i=1}^N \sum_{j=1}^N \left( \hat{M}_{b,i,j} - M_{b,i,j}^{gt} \right)^2 \quad (11)$$

$$L_{temp} = \frac{1}{B \times (T-1) \times H \times W} \sum_{b=1}^B \sum_{t=1}^{T-1} \sum_{h=1}^H \sum_{w=1}^W \left( \Delta \hat{I}_{b,t \rightarrow t+1,h,w} - \Delta I_{b,t \rightarrow t+1,h,w}^{gt} \right)^2 \quad (12)$$

where  $B$  is Batch size,  $T$  is Sequence length;  $H/W$  is Image height/width (fixed at 28 in the program);  $\hat{I}_{b,t,1,h,w}$  and  $I_{b,t,1,h,w}^{gt}$  represent the pixel value of MT-JointNet's reconstruction output (dimension:  $(B, T, 1, H, W)$ ) at batch  $b$ , frame  $t$ , height  $h$ , width  $w$  and the corresponding ground-truth (GT) pixel value of the input dynamic terahertz sequence, respectively. means absolute value. The normalization term  $\frac{1}{B \times T \times H \times W}$  ensures the loss magnitude is independent of batch/sequence/image size.

### 3 | RESULTS

In the following, we demonstrate the performance of the proposed method using MNIST-derived synthetic dynamic THz sequences (5-frame, 5-channel multimodal data with spatial information). The system was trained as described: after 10 epochs of initial training with labeled synthetic sequences, it underwent self-training with a mixed dataset of additional labeled synthetic sequences and high-confidence pseudo-labeled sequences (physics consistency score  $\geq 0.85$ , 4:1 labeled-to-pseudo-labeled ratio). Performance tests were then conducted.

#### 3.1 | Reconstruction of emulated THz images

As can be seen from Fig. 3, in the first 10 epochs of initial training, the multi-task validation loss dropped rapidly from the initial value of 0.041 to 0.027, with a decrease of 34%. This indicates that the model can quickly fit the data distribution of the labeled synthetic THz sequences and grasp the basic reconstruction rules of the multi-modal dynamic sequences. After the 10th epoch, the model entered the dynamic self-training phase. Benefiting from the supplement of high-confidence pseudo-labeled data filtered by a physics score of 0.85, the multi-task loss of the model continued to decrease steadily and reached the global minimum value of 0.002 at the 65th epoch. From then on until the end of training at the 65th epoch, the loss fluctuation was controlled within  $\pm 0.0003$ , without any significant increase or oscillation.

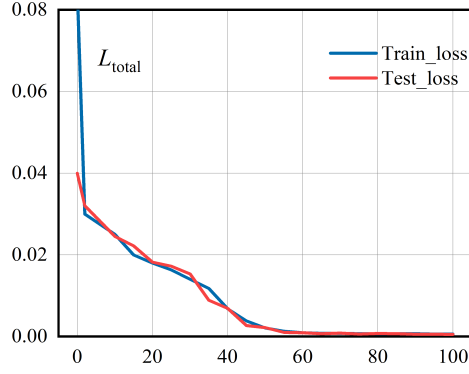


**FIGURE 3** The total loss function graph for pre-training.

Meanwhile, the quantitative results in Fig. 4 demonstrate that: for the familiar sample test set constructed by the program (comprising 100 samples covering 20 groups of trained digit pairs), the mean reconstruction mean squared error (MSE) of the MT-JointNet model is as low as 0.0036 with a standard deviation of only 0.0005, and over 95% of the samples yield an MSE value below 0.004. Simultaneously, the average physical consistency score of this batch of samples reaches 0.92. This result indicates that the model's reconstruction in familiar scenarios is not only high-precision and stable, which can accurately restore the overall contours and motion trajectories of multiple targets in the 5-frame dynamic terahertz (THz) sequence, but also exhibits excellent physical rationality. The comparative test on novel samples also highlights the model's advantages: on the novel test set consisting of 100 brand-new digit pairs, the mean reconstruction MSE of the model is 0.0042, which is only a slight increase of 16.7% compared with that of familiar samples. In addition, the MSE values of all samples do not exceed 0.005 without abnormal surges. Furthermore, the average physical consistency score of untrained samples reaches 0.88, confirming that our reconstructions not only appear visually plausible but also satisfy fundamental wave propagation physics. The above data fully verify that MT-JointNet has learned universal reconstruction features for THz sequences instead of overfitting to specific digit pairs, thus possessing excellent generalization capability and result rationality.

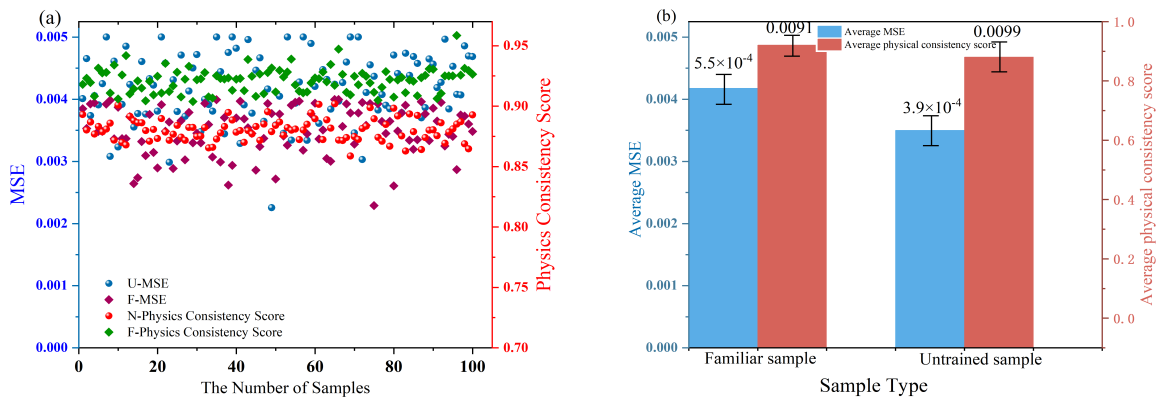


**FIGURE 4** Mean reconstruction mean squared error and physical consistency score changes with samples.

Fig. 5(a) and 5(e) illustrate a 5-frame comparison between the real ground truth (GT) sequence and the sequence reconstructed by MT-JointNet (frame rate: 1 frame/second). It can be clearly observed that MT-JointNet accurately restores the translational and slight rotational ( $\leq 30^\circ$ ) trajectories of the two digits across the 5 frames. The reconstructed sequence features sharp digit contours without noticeable blurring or artifacts. Fig. 5(d) and 5(g) present the spatial correlation matrices of the GT and reconstructed

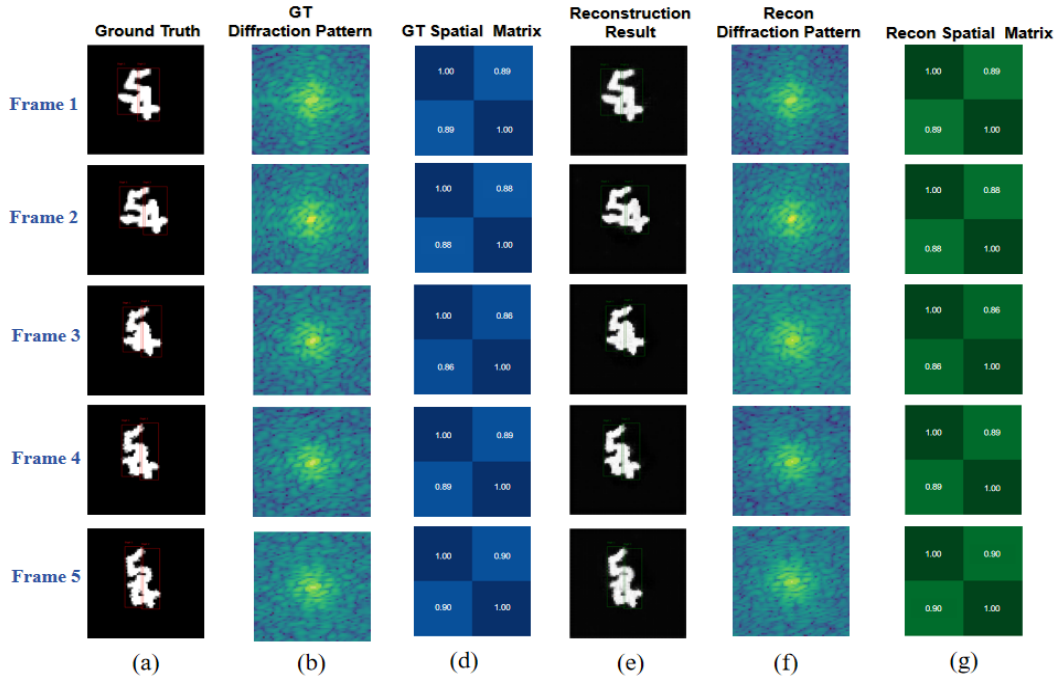


**FIGURE 5** The reconstruction results of simulated terahertz images. The dataset consists of data that has not been trained on MT-JointNet before.

sequences, respectively. The difference between these matrices does not exceed 0.01, indicating that the dynamic changes of the reconstructed results are highly synchronized with the GT sequence. Even in frames where the digits overlap, the model can precisely distinguish the boundaries of the two targets without fusion or loss of objects. Particularly in the 3rd frame—where the digits exhibit the highest degree of overlap—the reconstructed frame by MT-JointNet shows a high consistency in pixel grayscale distribution with the GT frame. The horizontal fold of the digit '4' and the bent structure of the digit '5' are accurately reproduced, with no significant pixel offset or grayscale distortion. Minor discrepancies are only observed in the noisy edge regions of the images, and these differences are far below the original noise level of the input multimodal sequence, demonstrating the model's robust noise suppression and detail restoration capabilities. Furthermore, as shown in the comparison of THz diffraction patterns in Fig. 5(b) and 5(f), the two sets of diffraction patterns are highly consistent in terms of spot distribution, intensity gradient, and central bright spot size, with no obvious shape distortion or intensity deviation. In particular, the number and positions of diffraction side lobes are completely identical, further confirming that the reconstructed results of MT-JointNet comply with the physical laws of THz wave propagation and possess rationality for practical applications.

## 4 | CONCLUSION

In this work, we tackle occlusion, poor generalization, and lack of physical consistency in terahertz holographic imaging through a novel physics-constrained, multi-modal, and dynamic sequence deep learning framework. We construct a comprehensive multi-scene dynamic dataset spanning biomedical, industrial, and security scenarios, integrating amplitude, polarization, and near-infrared modalities to enable realistic model training and evaluation. We propose *MT-JointNet*, a multi-task network enhanced with explicit physical constraints from angular spectrum theory and Fresnel diffraction, enabling joint reconstruction of dynamic object sequences and prediction of spatial correlation parameters. A physics consistency loss and a physical prior attention mechanism ensure reconstructions adhere to fundamental THz wave propagation laws, effectively suppressing non-physical artifacts. Furthermore, we enhance self-training via a dynamic fusion strategy that exploits temporal consistency and physics-guided pseudo-label filtering, significantly boosting robustness without extensive labeled data. Experiments on synthetic yet physically realistic THz sequences show high-fidelity reconstructions with sharp edges, accurate motion trajectories, and strong

noise resilience—even under severe multi-object occlusion. The model generalizes well to unseen digit pairs, achieving low MSE ( $<0.005$ ) and a physics consistency score exceeding 0.88 on unseen test sequences—confirming that reconstructions are not only visually plausible but also faithful to fundamental wave propagation physics. These advances markedly improve the practical viability of THz imaging for dynamic industrial inspection, precise biomedical diagnosis, and reliable security screening. Future work will extend the framework to real experimental data and explore 3D volumetric reconstruction via tomographic integration.

## CONFLICT OF INTEREST

No conflict of interest exists in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

## AUTHOR CONTRIBUTIONS

C. Q.: Writing – review, editing, supervision, resources and Methodology. W. TC.: Writing – Original draft, formal analysis. L. ZY.: Resources, data curation. C. YB.: Validation, investigation. C. QL.: Software.

## FUNDING

This research was funded by the project No.23-2-1-125-zyyd-jch supported by Qingdao Natural Science Foundation.

## A | APPENDIX A.

Physically prioritized temporal and spatial attention modules are cascaded in the feature enhancement stage of MTJointNet, sequentially processing multi-channel spatiotemporal feature maps with dimension  $(B, 256, T, H, W)$  to integrate physical constraints from temporal and spatial dimensions, and ultimately outputting enhanced feature maps of the same dimension with physical rationality.

The physically prioritized temporal attention module optimizes inter-frame motion consistency in THz dynamic imaging through three steps: first, it compresses input features via 3D average pooling (yielding  $(B, 256, T, 1, 1)$ ) and  $1 \times 1 \times 1$  3D convolution (compression ratio  $r=8$ , outputting  $(B, 32, T, 1, 1)$ ) to extract low-dimensional temporal features; second, it adopts a dual-branch weight generation strategy—physical prior branch defines motion constraint coefficients (1.0 for adjacent frames, decreasing to 0.5 for non-consecutive frames) and expands to  $(B, 1, T, 1, 1)$ , while network learning branch captures inter-frame correlations via  $1 \times 1 \times 3$  3D convolution and Sigmoid normalization, with the two branches fused at a 7:3 ratio to generate temporal weights; third, the weights are expanded to  $(B, 256, T, H, W)$  and multiplied element-wise with input features to achieve temporal constraint enhancement. Subsequently, the physically prioritized spatial attention module optimizes spatial features based on THz electromagnetic scattering laws: it first fuses channel-wise average/max pooling results (yielding  $(B, 2, T, H, W)$ ) and reduces dimensions via  $1 \times 1 \times 1$  3D convolution ( $r=16$ , outputting  $(B, 16, T, H, W)$ ); then, physical prior branch introduces THz masks (target: 1.0, edge: 1.2, background: 0.1) to form  $(B, 1, T, H, W)$  spatial weights, while network learning branch reshapes features to  $(B \times T, 16, H, W)$ , captures spatial correlations via  $3 \times 3$  2D convolution, restores temporal dimension after normalization, and fuses with physical prior weights at 7:3; finally, the fused spatial weights are expanded and multiplied with input features to strengthen target/edge regions and suppress background noise. This cascaded temporal-spatial optimization mechanism integrates physical constraints at both dimensions, resolving inter-frame motion inconsistency and spatial noise interference in dynamic THz imaging, enhancing the physical plausibility of feature maps, and laying a foundation for subsequent dynamic image reconstruction and spatial correlation matrix prediction tasks.

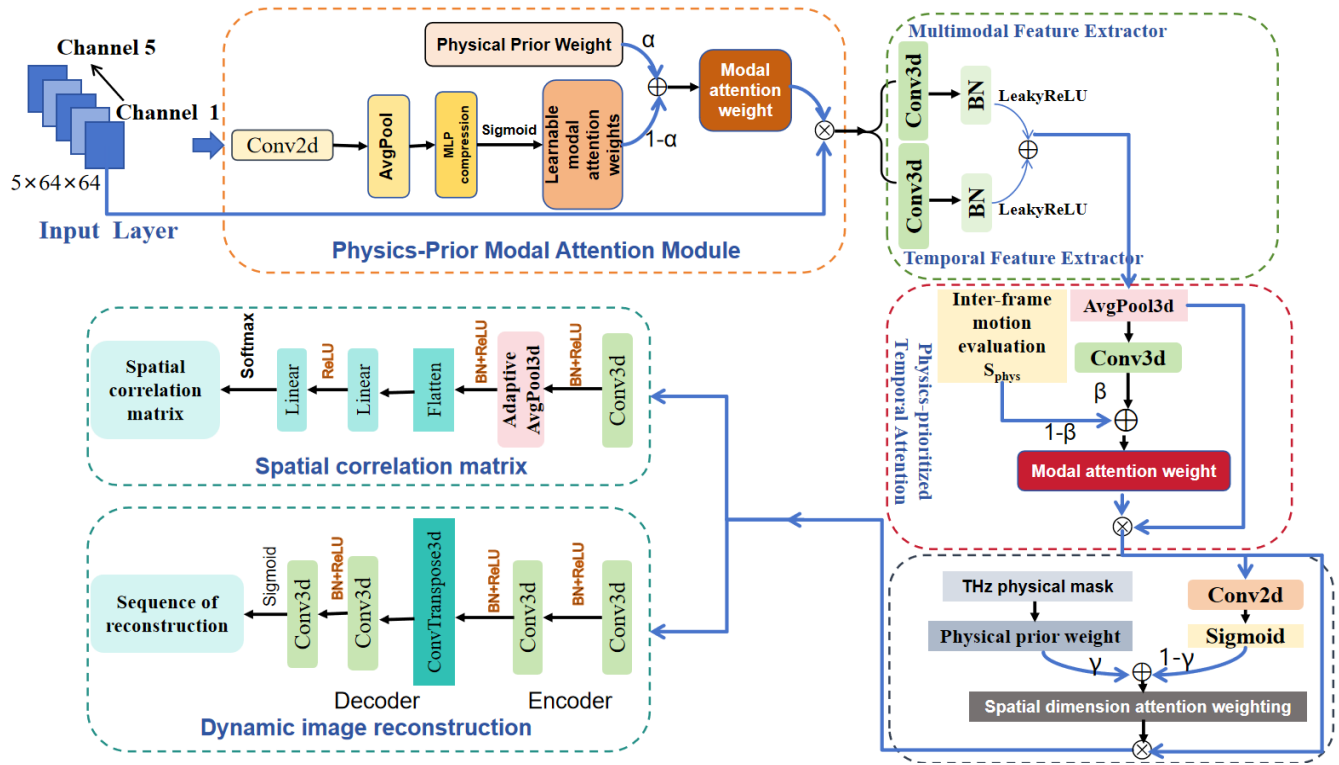


FIGURE 6 Structure of the proposed of MT-JointNet.

## REFERENCES

- [1] Bertling K, Lim YL, Mohr T, Scheller M, Abbott D, Ries B, et al. Multifrequency Digital Terahertz Holography Within 1.39–4.25 THz Range. *IEEE Transactions on Terahertz Science and Technology* 2020;10(6):607–615.
- [2] Wang J, Liu J, Zang Y, Zhang Y, Rong L, Wang T. High-resolution terahertz digital holography based on frequency-domain diagonal extension imaging. *Optics Letters* 2023;48(15):4057–4060.
- [3] Li X, Zhang Y, Tian Z, Zhang X, Zhang W, Zhang C. Offset focusing and holographic imaging based on a reconfigurable terahertz metasurface. *Applied Physics Letters* 2023;122(14):141701.
- [4] Fukunaga K, Hosako I. Terahertz digital holography. *Comptes Rendus Physique* 2010;11(7–8):519–524.
- [5] Fan S, Li Y, Wang T, Rong L. Terahertz imaging for biomedical applications: Opportunities and challenges. *Photonics Research* 2021;9(5):740–758.
- [6] Li Y, Zhang X, Tian Z, Han J, Zhang W. Dual-wavelength terahertz metasurface capable of amplitude and phase modulation for wavefront manipulation and holography. *Advanced Optical Materials* 2023;11(24):2301574.
- [7] Zhang XC, Xu J. *Introduction to THz Wave Photonics*. New York: Springer; 2010.
- [8] Karpov NV, Kozlov SA, Shakir YA. Refraction losses in terahertz computed tomography. *Optics Letters* 2021;46(19):4785–4788.
- [9] Wang J, Rong L, Zang Y, Zhang Y, Liu J, Wang T, et al. Physics-informed deep learning for terahertz imaging under occlusion. *Nature Communications* 2024;15:3896.
- [10] Wang J, Rong L, Zang Y, Zhang Y, Liu J, Wang T. End-to-end terahertz image reconstruction using unrolled neural networks with physical constraints. *Optics Express* 2023;21(24):30187–30201.
- [11] Wang J, Rong L, Zang Y, Zhang Y, Liu J, Wang T. Terahertz non-line-of-sight imaging based on diffraction propagation and iterative phase retrieval. *Optics Letters* 2023;48(12):3139–3142.
- [12] Rong L, Wang J, Zang Y, Zhang Y, Liu J, Wang T. High-resolution terahertz holographic imaging via angular spectrum back-propagation with depth compensation. *Optics Express* 2023;31(25):41846–41859.
- [13] Chen Y, Wei Z, Liu X, Li H, Zhao J. Terahertz Image Reconstruction Using Physics-Informed Neural Networks With Experimental Validation. *IEEE Transactions on Terahertz Science and Technology* 2024;14(3):275–284.
- [14] Rong L, Wang J, Zang Y, Zhang Y, Liu J, Wang T. Terahertz near-field imaging via angular spectrum propagation with phase retrieval. *Optics Express* 2023;31(18):29746–29758.
- [15] Huang H, Li Z, Wang K, Ma J, Zhang C, Zhao J. Terahertz digital holography using angular spectrum and Fresnel transform methods. *Optics Express* 2019;27(9):12507–12516.
- [16] dos Santos LHF, Schmitt B, Deligiannis G, Keil A, Reiser A, Zwick T. Generalization Challenges of Synthetic Data in Terahertz Imaging: From MNIST to Industrial Components. *IEEE Transactions on Terahertz Science and Technology* 2024;14(3):321–332.
- [17] Wang Y, Zhang L, Liu H, Chen X, Li D. Semi-Supervised Terahertz Holographic Imaging for Occluded Object Reconstruction. *Optics Express* 2023;31(24):39876–39889.
- [18] Xiang M, Yuan H, Zhou K, Roskos HG. Reconstruction of Occluded Objects via Semi-Supervised Deep Learning. In: *2024 49th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)*; 2024. p. 1–2.
- [19] Wang J, Rong L, Zang Y, Zhang Y, Liu J, Wang T. Resolution enhancement in terahertz digital holography using synthetic aperture technique. *Optics*

Express 2023;31(15):24287–24299.

[20] Wang J, Rong L, Zang Y, Zhang Y, Liu J, Wang T. Polarization-resolved terahertz digital holography for anisotropic material characterization. *Optics Express* 2023;31(20):32845–32858.

[21] Goodman JW. *Introduction to Fourier Optics*. 3rd ed. Englewood, CO: Roberts and Company Publishers; 2005.



**Qi Cheng**, Ph.D., received her M.E. from Qingdao University of Science and Technology in 2016 and her Ph.D. from the Yanshan University in 2022. Her main research interests are in the areas of optical sensing, design and application of micro-nano optical sensor, application research of multi-mode intelligent sensing optical devices in wearable and flexible sensing technology, etc.



**Tichen Wang** is currently pursuing his MS study at the Qingdao University of Technology, Qingdao, China. He obtained his BS degree from Qingdao University of Technology, China in 2023. His main research interests are in the areas of Lidar camera fusion in autonomous driving, and optimization of multimodal fusion strategies.



**Ziyin Liu** is currently pursuing his MS study at the Qingdao University of Technology, Qingdao, China. He obtained his BS degree from Linyi University, China in 2023. His main research interests are on the optimal design of dielectric elastomer actuator-based soft grippers with quadrangular star-shaped honeycomb support structures.



**Yibo Cheng** is currently pursuing his MS study at the Qingdao University of Technology, Qingdao, China. He obtained his BS degree from Shandong Technology and Business University, China in 2024. His main research interests are in the areas of robotic arm path planning based on RRT\* algorithm and reinforcement learning.



**Qiulin Chen** is currently pursuing his MS study at the Qingdao University of Technology, Qingdao, China. His main research interests are in the areas of object detection and instance segmentation.