

Received 15 July 2024, accepted 31 July 2024, date of publication 6 August 2024, date of current version 16 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3439358

## RESEARCH ARTICLE

# A 3D Reconstruction of Terahertz Images Based on the FCTMVSNet Algorithm

XIAOJIN WU<sup>1</sup>, HAIXIAN LIU<sup>2</sup>, FAN BAI<sup>3</sup>, XUDONG LU<sup>3</sup>, YUAN GAO<sup>1</sup>, AND LUN LI<sup>1</sup> 

<sup>1</sup>Institute of Machinery and Automation, Weifang University, Weifang 261061, China

<sup>2</sup>Department of Respiratory and Critical Care Medicine, Weifang City People's Hospital, Weifang 261044, China

<sup>3</sup>College of Equipment Engineering, Shenyang Ligong University, Shenyang 110159, China

Corresponding authors: Lun Li (11408907652@163.com) and Fan Bai (snow-wind-001@163.com)


This work was supported in part by the Natural Science Foundation of Shandong Province under Grant ZR2020KF033, Grant ZR2023MF047, and Grant ZR2021QB210; in part by Weifang High-Tech Zone Science and Technology Benefit People Program under Grant 2021KJHM55; in part by the Doctoral Research Start-Up Fund Project of Weifang University under Grant 2022BS22; and in part by the Project of Science and Technology on Electromechanical Dynamic Control Laboratory, China, under Grant 6142601220603.

**ABSTRACT** The terahertz range, as a type of electromagnetic wave with wavelengths between microwaves and the infrared band, has the characteristics of penetration, low energy and a stable absorption spectrum of specific substances, and is widely used in non-destructive testing, human security inspections, biological tissue diagnoses and military detection. In particular, terahertz wave 3D imaging technology can detect the internal information of the target of detection, and it has become the focus of current research. This study carried out research on 3D reconstruction and object detection algorithms based on terahertz images. In view of the problem that the MVS (Multi-ViewStereo) series of 3D reconstruction algorithms ignore the context information between the cost layers and have unsatisfactory reconstruction effects when used on complex regions, an improved MVSNet 3D reconstruction algorithm FCTMVSNet (Feature and Cost Transformer Depth Inference for Unstructured Multi-view Stereo) based on Transformer is proposed here. A structured object recognition algorithm was designed to provide theoretical support for subsequent terahertz image-based object detection algorithms.

**INDEX TERMS** Terahertz imaging, transmission type, FCTMVSNet, three-dimensional reconstruction.

## I. INTRODUCTION

Terahertz waves have low energy. The energy of light waves in the terahertz frequency level is only a few electron volts, and 1 electron volt is equal to the amount of electron charge in an element, which is about Joule, so they will not damage the object to be detected. Terahertz light waves can penetrate most non-metallic materials, such as ceramics, plastics, foam and nylon, to detect hidden objects. Therefore, it can replace traditional X-ray detection methods and be used for security detection in public areas such as airports, stations and subways to detect dangerous items such as knives, explosives and guns and ammunition. Terahertz-based human security technology can detect objects hidden under clothes, but a single-view terahertz detection system is limited by the shooting angle, and some angles cannot obtain the complete feature

The associate editor coordinating the review of this manuscript and approving it for publication was Abedalrhmman Alkhateeb .

information of the object, and thus the nature of the object cannot be accurately judged. Therefore, terahertz detection methods based on multi-view motion reconstruction is worth studying [1], [2], [3], [4].

Because terahertz three-dimensional imaging technology can better obtain the internal information of the sample, it has become a research hotspot. Three-dimensional Terahertz imaging technologies mainly include terahertz computed tomography (CT) imaging, terahertz diffraction tomography, terahertz tomography and terahertz digital holography. Buma and Zhang [5], combined the synthetic aperture focusing technique with the point-by-point imaging technique to construct a 3D image of the target [6]. In addition, they used a weighted sum algorithm to solve the problem of sidelobe artifacts in the reconstructed image. Abraham et al. investigated the effect [6] of objects with a large refractive index on 3D tomography using terahertz pulse imaging. In 2011, the University of Electronic Science and Technology of

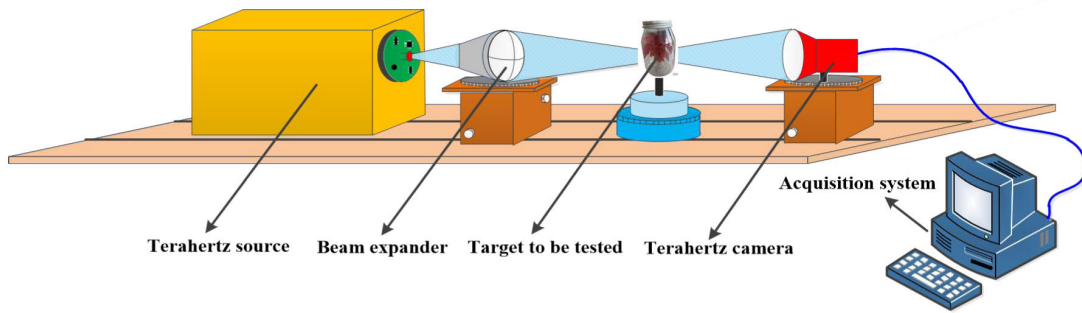


FIGURE 1. Terahertz imaging principle.

China designed a continuous THz tomography system with a planar array detector using CO<sub>2</sub>-pumped CH<sub>3</sub>OH to generate THz waves [7]. In 2012, a 0.34 THz superheterodyne 3D imaging radar system was designed by the University of St. Andrew's, which was used for display and processing by transmitting co-polarization, receiving co-polarization and cross-polarization. The project started in 2008, and through continuous improvement, an imaging frame rate of 10 frames/s at 20 m was achieved [8], [9]. In 2016, Tripathi et al. performed CT imaging of plastic objects with a narrow linewidth, a tunable terahertz parameter source and terahertz conversion frequency upconversion detection technology. Because this method converts the terahertz frequency detection to near-infrared band detection, the dynamic range of detection could reach 90 dB near 1.5 THz. It could reflect the internal information of the object and the location of the defects well [10]. In 2017, Zhou et al. successfully reconstructed 3D images of ceramic samples by using a Uni-Traveling-Carrier Photo Diode (UTC-PD) to generate 90-140 GHz of low-coherence terahertz radiation. These studies showed that terahertz waves have great practical value in non-destructive testing [11].

On the basis of analyzing the characteristics of terahertz imaging, this article proposes an improved MVSNet 3D reconstruction algorithm based on Transformer (FCTMVSNet, feature and cost transformer depth inference for unstructured multi-view stereo) to address the issue of using convolution as a feature extraction network for 3D reconstruction algorithms such as MVSNet [12] and PA-MVSNet [13] in the MVS series, which ignore contextual information between the cost layers and do not achieve ideal reconstruction results in complex areas. The traditional convolutional feature extraction network is replaced by the self-attention mechanism to solve the problem that the feature extraction network of the traditional 3D reconstruction algorithm is limited by the spatial location information and is insensitive to global information. At the same time, the inter-layer attention mechanism of the cost body is proposed to improve the accuracy of the network, which provides a theoretical basis for the engineering applications of the algorithm. The structural position feature was used to slice and output the obtained 3D model, and specific object

detection was realized. At the same time, the reconstruction accuracy of this method has been significantly enhanced, particularly in complex regions and especially when dealing with fuzzy textures such as terahertz images. Meanwhile, in this paper, the combination of FCTMVSNet and FCOD is proposed to implement the comprehensive recognition algorithm for multi-view terahertz images and achieve accurate recognition of objects with low terahertz resolution.

## II. DESIGN OF THE 3D TERAHERTZ IMAGING SCHEME

### A. TERAHERTZ IMAGING PRINCIPLE

In this paper, a multi-view stereo matching algorithm is employed to accomplish 3D reconstruction of terahertz targets. This requires multiple cameras to conduct multi-view imaging around the target. Nevertheless, due to the constraint of the imaging equipment, the terahertz source and the terahertz receiver are unable to achieve multi-view acquisition simultaneously. Thus, the target object is rotated to achieve multi-view imaging during the experiment.

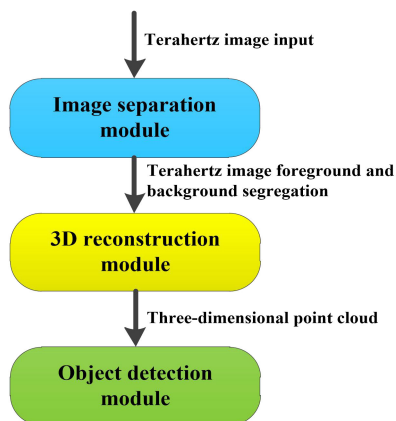
The terahertz three-dimensional imaging system is composed of a terahertz source, a terahertz beam expanding mirror, a terahertz camera, a measured object, and a rotating mechanism of the object, as depicted in Figure 1. The terahertz source transmits the terahertz wave to focus on the central axis of the target through the terahertz beam expanding mirror and then converges it to the terahertz detector to achieve single point image acquisition.

### B. TERAHERTZ 3D IMAGING ALGORITHM DESIGN

Terahertz imaging is affected by the imaging device, the definition of the collected terahertz image can be low, and there are often diffraction patterns in the image. At the same time, in an active terahertz imaging system, when the terahertz light source passes through the object, a shadow area overlapping the foreground and background will be formed. The overlapping shadow area will mask the information on the surface texture of the object, and effective feature matching cannot be carried out during 3D reconstruction, which affects the quality of the reconstruction.

The direct use of terahertz images for target recognition and detection will be affected by factors such as perspective and the image's clarity. Effective pre-processing of terahertz

images is required to make full use of the multi-view 3D information for reconstruction and recognition of a target. Therefore, the overall scheme of this study is shown in Figure 2. The overall scheme is composed of three modules: the image separation module, the 3D reconstruction module and the target detection module.



**FIGURE 2.** Schematic of the design of the target detection algorithm based on 3D terahertz imaging.

As terahertz imaging utilizes the electromagnetic radiation characteristics of the terahertz band, it can penetrate many non-metallic materials (such as plastics, paper, fabrics, etc.) and has certain transmission and reflection capabilities for many common non-conductive substances and biological tissues. Therefore, terahertz imaging is also known as terahertz transmission imaging (plus a terahertz transmission image).

Due to the superposition of the foreground and background textures in the terahertz transmission image, a feature matching error is generated, and 3D reconstruction of the object cannot be performed. Therefore, the terahertz image transmission image needs to be separated from the foreground and background before reconstruction, so that the terahertz image can meet the standard for reconstruction. Secondly, on the basis of the traditional 3D reconstruction algorithm MVSNet, the Transformer is used to replace the convolutional network, so that the network can obtain the related information between images and retain the feature-related information between the cost layers. Finally, the terahertz point cloud images obtained by structuring multiple fixed angles can completely express the feature information of the object, thus improving the detection ability of the network for objects.

### III. TERTZ IMAGE 3D RECONSTRUCTION AND TARGET DETECTION ALGORITHM

#### A. OVERALL STRUCTURE OF FCTMVSNET NETWORK

Aiming to solve the problem that the MVS series of 3D reconstruction algorithms ignore the context information between the cost layers and their reconstruction effect is not ideal for complex areas, this study proposed an improved 3D reconstruction algorithm known as FCTMVSNet, which is based

on Transformer, and designed a view-structured target recognition algorithm to realize the 3D reconstruction of targets based on terahertz images.

The overall network structure of FCTMVSNet is shown in Figure 3. The entire network is mainly composed of the feature extraction module, the cost loss body construction module, the cost interlayer attention mechanism module, the depth map estimation module and the depth map optimization fusion module.

Due to the immaturity of terahertz imaging devices, a terahertz source with an output power of 150mW @ 2.52THz (118.8 $\mu$ m) from the Chengdu Precision Optical Engineering Research Center of the Chinese Academy of Nuclear Physics was selected for the experimental system built in this article, and the resolution of the Huirui photoelectric terahertz camera used for collecting the terahertz images was 640  $\times$  480. Due to limitations in the optical path and the terahertz source's frequency, the terahertz dataset captured here was insufficient to support the subsequent experiments, the experimental system built and the terahertz image data collected are shown in Figure 4. To verify the feasibility of our algorithm for terahertz image reconstruction, we used the terahertz dataset captured by Hongke Electronics Technology and the nanoelectronics team from the IMS Laboratory of the University of Bordeaux in France [18]. The equipment they used was a TeraCascade 1000 high-power terahertz source (1.3 mW, 2.5 THz output), a dual-mirror galvanometer, a 45° off-axis parabolic reflector, a INO microcalorimeter array(288  $\times$  384 pixels) and a TeraLens (F/0.8) lens. Due to the particularity of terahertz imaging technology, there was a transmissive imaging effect. If a 2.5 THz frequency terahertz source was used, a resolution of 250  $\mu$ m could be achieved. This was sufficient to verify the effectiveness of the algorithm proposed in this study.

Firstly, the multi-view terahertz transmission-separated image was processed through Transformer's feature extractor to obtain the corresponding feature vectors. The homography matrix projected the remaining view feature vectors and corresponding camera parameters onto the main view to form the volume of the cost loss. Then a self-attention mechanism was used for each cost body layer to extract the contextual information. Finally, the initial depth map of the main view was obtained through the encoder and decoder layers.

#### B. FEATURE EXTRACTION AND 4D COST LOSS VOLUME

In traditional 3D reconstruction algorithms, the convolution operation uses two important spatial constraints, namely the weight-sharing mechanism and the translation invariance of the features extracted by convolution layers, for learning and extracting visual features. However, this constraint makes the convolution have poor ability to perceive the global position of the features and it only cares about whether these decisive features exist. Due to the nature of the convolution operators, the feature map of convolution has local sensitivity, that is, each convolution operation only considers the local information of the original data [19], [20], [21], [22].

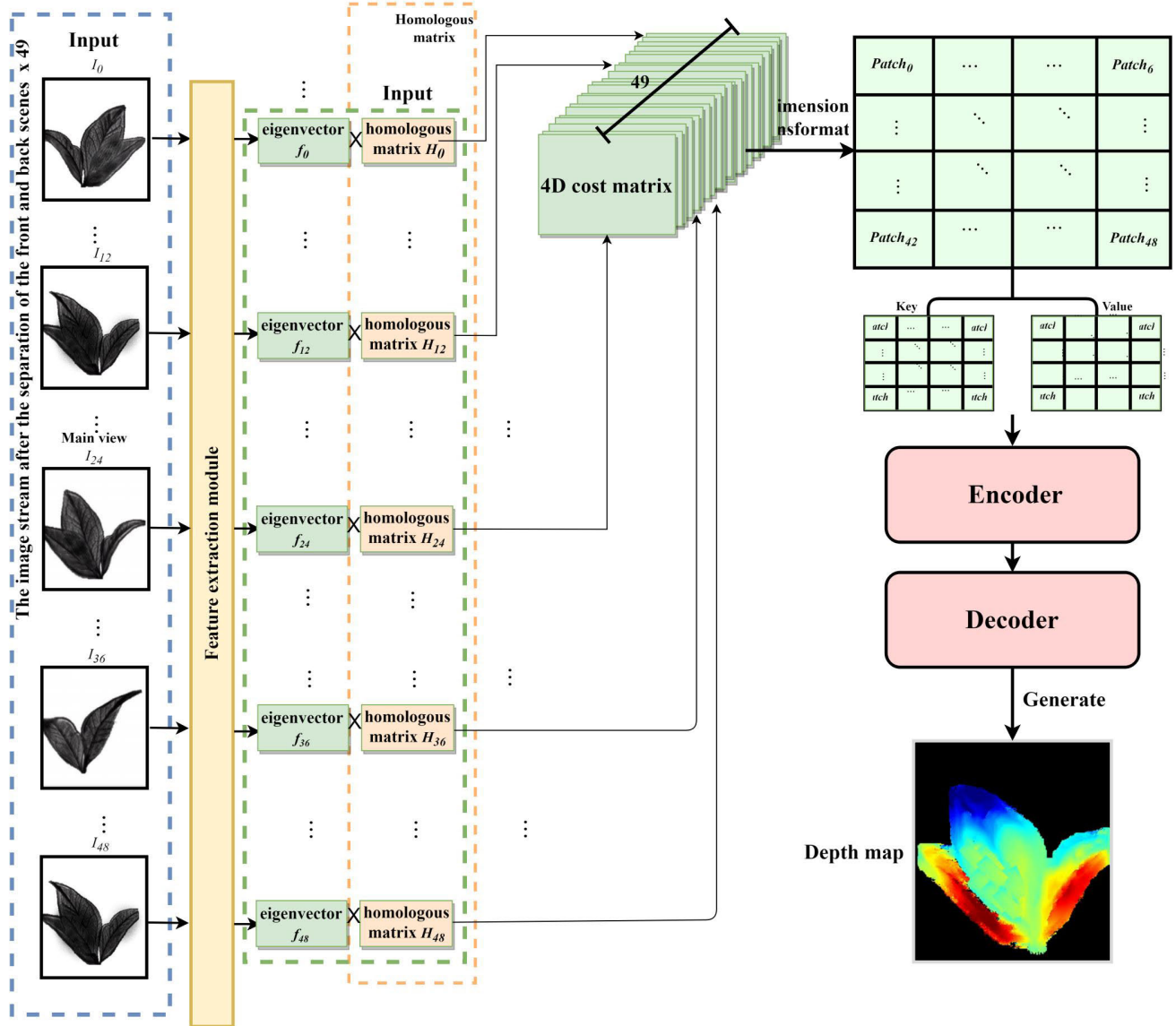


FIGURE 3. Diagram of the network structure of the 3D reconstruction algorithm based on multi-view stereo depth of field inference (FCTMVSNet).

These reasons lead to the inductive bias of CNN, which lacks an overall grasp of the input data itself and can only extract effective local information, but cannot extract the long-distance features of the global data [23].

The feature maps extracted by Transformer using its unique self-attention mechanism are not limited by spatial information, as in the case of convolution. On the contrary, it can effectively learn the target area and background information, as well as the correlation between images, as shown in Figure 5. In this study, when constructing the network of the 3D reconstruction algorithm, the first few layers of the network used the self-attention mechanism to extract the features from the terahertz images.

According to the conventions of information retrieval, the features are grouped into Q values, K values and V values.

The Q values retrieve relevant information from the V values based on the attention weight obtained by the dot product of Q and K corresponding to each V. The form of the attention layer is  $\text{Attention}(Q,K,V)=\text{softmax}(QK^AT)$ .

The attention mechanism measures the similarity of the features between Q and K, and retrieves the information from V based on the calculated weights.

We constructed the 4D cost loss volume by using the camera's parameters and the feature maps extracted by Transformer. After deriving the depth map, a cost loss volume was constructed on the basis of the conical principle of using the first input photo as the main perspective of the camera. In 3D reconstruction, the perspective cone principle is used to describe the representation of the camera's or observer's field of view in 3D space. It is based on the principle of



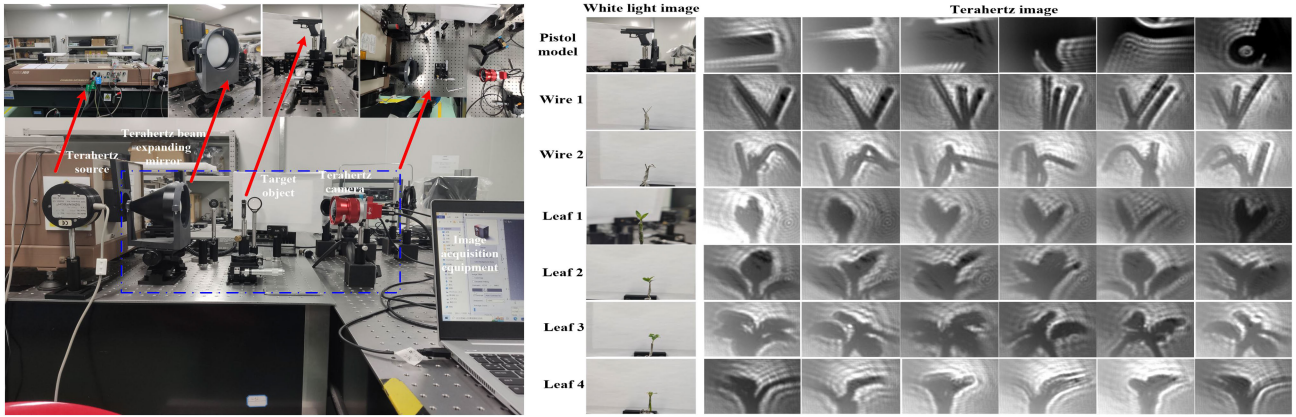


FIGURE 4. The terahertz experimental system built by this team and the collected terahertz experimental images: (a) Experimental system used for 3D terahertz imaging. (b) Experimental results of terahertz imaging.

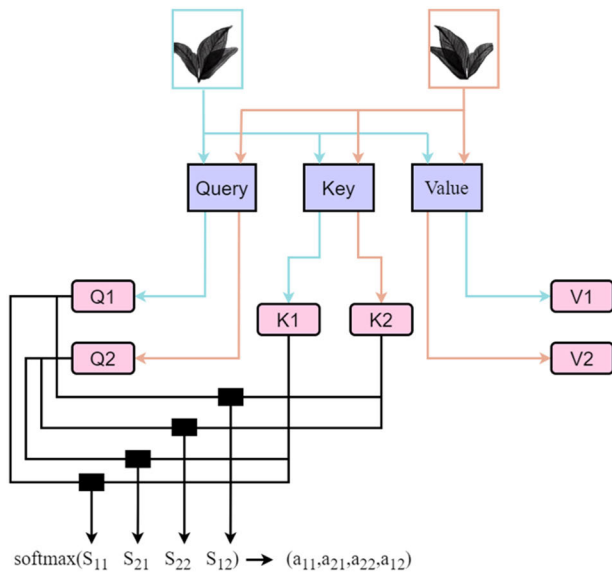


FIGURE 5. Attention mechanism.

perspective projection, which limits the field of view to a conical area, where the camera’s position is the top of the cone, and the distance and breadth of the field of view are determined by the height of the cone and the width of the bottom. The main perspective image and other perspective images are extracted through Transformer’s features to obtain the corresponding feature vectors. Due to the differences in the images from each perspective, due to their computational complexity, depth information cannot be extracted from all views. Therefore, depth information can only be extracted from the images from the main perspective to obtain a depth map. Therefore, images from other perspectives need to be transformed to the perspective of the main perspective image through homography.

All other viewpoint feature maps are transformed into the stereo space corresponding to the main viewpoint image through a homography transformation. The homography

transformation can be performed using Formula (1),

$$X \sim H_i(d) \cdot X \tag{1}$$

$$H_i(d) = K_i D_i \left( I - \frac{t_2 - t_1}{d} \cdot R_1^T \cdot R_2^T \right) \tag{2}$$

where “ $\sim$ ” represents depth-equivalent mapping;  $H_i(d)$  represents the depth of the other viewpoint images mapped to the cone space corresponding to the main viewpoint image;  $I$  represents the reference image;  $K$ ,  $R$  and  $t$  represent the camera’s internal and external parameters and horizontal displacement corresponding to the feature map; and  $d$  represents the depth information. Deep equivalence mapping is a technique used to map depth information from one perspective to another. In 3D reconstruction and multi-view stereo vision, when there are depth maps from multiple perspectives, depth equivalent mapping can convert these depth maps into a shared depth map for consistent 3D geometric calculations and rendering.

The feature maps of a certain number of the remaining viewpoints are mapped to form a feature aggregation  $\{V_i\}_{i=1}^N$ , and then multiple feature aggregations  $\{V_i\}_{i=1}^N$  are merged into a cost loss body  $c$ . In order to adapt to any number of viewpoint inputs, a variance-based method of calculating the loss body was adopted in this study.

$$V = \frac{W}{4} \cdot \frac{H}{4} \cdot D \cdot F \tag{3}$$

where  $W$ ,  $H$ ,  $D$  and  $F$  are the width, height, the number of visual maps and the number of channels, respectively; and  $V$  represents the volume of the feature map.

The mapping relationship of loss volume is as follows.

$$c = M(V_1, \dots, V_N) = \frac{\sum_{i=1}^N (V_i - \bar{V}_i)^2}{N} \tag{4}$$

where  $\bar{V}_i$  is the average of all feature aggregations.

In this study, the traditional calculation method was continued, and  $c$  was calculated by pairwise pairing of images from the main view angle and images from other views. However, in the process of calculation, the cost loss of the

main view angle should not be biased when choosing what to focus on, and each view angle should be treated equally. The traditional calculation method is to average the loss of multiple perspectives, which will not provide the network with the information about the differences in the features. Therefore, this study used variance to replace the average value.

### C. ESTIMATION AND FUSION OF THE DEPTH MAP

When the surface of the object is blocked and distorted, the resulting cost loss body usually carries some noise. To solve this problem, the depth map is constrained by smoothness to reduce the influence of noise. At the same time, a probabilistic aggregate  $P$  is generated on the basis of the cost loss body  $c$  for the inference of the depth map. With multi-scale 3D CNN being used to regularize the cost loss body, four cost bodies of different sizes were constructed to aggregate the surrounding information. At the same time, in order to further reduce the memory consumption, the cost loss body of 32 channels was reduced to eight channels after the first 3D convolution, and the images' size was reduced in the second and third layers. The final 3D convolution layer outputted a single channel polymer, which was then probabilistically unified using softmax operations in the depth direction. The result was a probabilistic aggregate, which stored the depth probability of the main view image in stereoscopic space [25-26].

The easiest way to obtain a depth map from a probabilistic aggregate is a preliminary estimate, which is made by calculating the expected value of each voxel depth estimate by Formula (5), taking the learnability of the network into account.

$$D = \sum_{d_{\min}}^{d_{\max}} d \times P(d) \quad (5)$$

where  $P(d)$  is the probability estimate of all voxels at depth  $d$ .

For each sample, we constructed a cost loss volume, and the maximum and minimum values for estimating depth were different, so we hoped to generate a continuous estimated value. The output depth map is shown in Figure 6a. Its size was the same as that of the 2D feature map, which was reduced by four times compared with the input image. The probability distribution along the depth direction can also reflect the quality of the depth map. Although multi-scale 3D CNN has a strong ability to regularize probability into a single mode of distribution, as shown in Figure 6b, this study noted that for pixels with matching errors, their probability distribution in the depth direction was relatively discrete and could not be concentrated, as shown in Figure 6c.

Based on this observation, we used the average depth  $\bar{d}$  to replace a small range of the depth probability. At the same time, the reconstructed boundary of the initial depth map was sometimes overly smooth due to the large receptive field during regularization, while the main view angle image in the natural scene contained boundary information. To solve this problem, the main view angle image was shrunk by

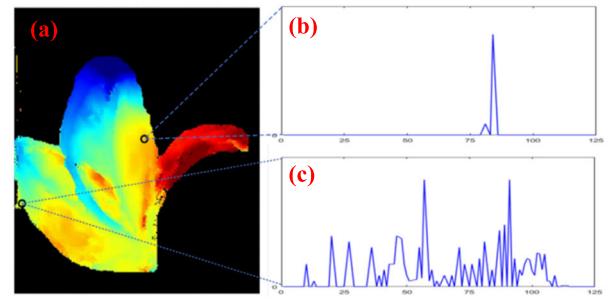


FIGURE 6. Initial depth map.

one-quarter during processing to make it the same size as the initial depth map. At the same time, the depth map was normalized to prevent deviation at a certain depth ratio. The two processed images were channel-stitched. Then the stitched four-channel image was put into a convolutional network with a four-layer residual structure for fusion of the information. Finally, the single-channel feature map outputted by the residual network was restored to the interval of the depth hypothesis (the reverse of the normalization process), and it was added element by element to the initial depth map, thus obtaining the optimized depth map.

The depth maps of different viewpoints were fused into a unified point cloud representation, and the visibility-based fusion algorithm was used to minimize the influence of occlusion, illumination and other factors, thus minimizing the depth occlusion and conflict between different viewpoints.

To further suppress reconstruction noise, the visible view of each pixel was determined in the filtering step, and the average of all the reprojected depths was used as the final depth estimate of the pixel. The fused depth map was projected directly into space to generate a 3D point cloud.

### D. OBJECT DETECTION ALGORITHM FOR THE STRUCTURED IMAGE OF THE POINT CLOUD

The terahertz images were reconstructed to form a point cloud, which was a 3D stereoscopic effect. Images taken from multi-structured fixed angles of interest could express all the feature information of the point cloud, effectively solving the problem of traditional object detection, which is limited by the low recognition accuracy of shooting angles. On this basis, the focus and CSP (center and scale prediction) based object detection algorithm (FCOD) was used as the target detector for the point cloud of structured terahertz images.

The network structure of FCOD consists of four parts: the feature extraction module, the feature fusion module, the feature enhancement module and the prediction module. The network's structure is shown in Figure 8.

The point cloud of the structured terahertz image was first extracted by the FCFE feature extraction module, which was processed by Focus and CSP, the core steps of the FCFE feature extraction module. As shown in Figure 9, Focus is similar to the sampling operation, which takes the value of every other pixel in each input image, so that one image



FIGURE 7. Schematic diagram of structured feature extraction.

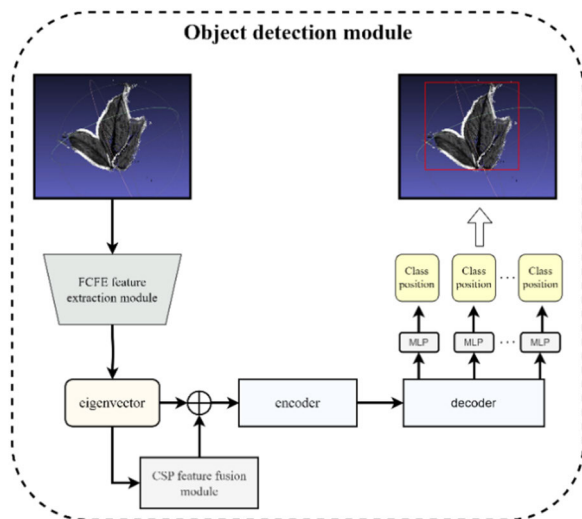


FIGURE 8. Network structure of the FCOD target detection algorithm.

will become four images, and the feature information of these four images is similar and complementary. This interval sampling operation does not cause the loss of the feature information of the original image, and the information on the width and height of the image will also be projected onto the channel space. Compared with the original image, the number of channels is expanded four times. Finally, the sampled image is extracted by the convolution layer, and the downsampling feature map of all information of the original image is obtained.

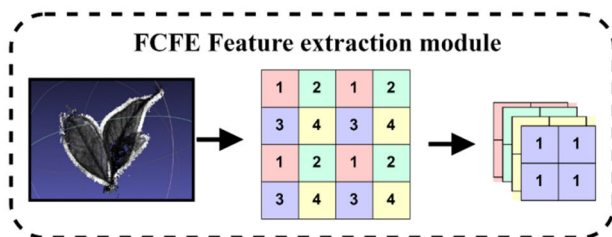


FIGURE 9. Focus slice processing.

The input image is sliced by the Focus module and then entered into the CSP module for deep feature extraction. As shown in Figure 10, the input feature map is divided into

two parts. One part is passed through the CBS layer (convolution, layer normalization and SiLU activation) and then through multiple residual layers, and finally through a convolution layer to extract the features. The other part is directly entered into the calculation of the convolution layer. The two parts are stitched, and then undergo BN and LeakyReLU layer normalization and activation, and finally pass through a CBS layer to extract the features. This can increase the depth of the network to extract more fine information on the features, and can also avoid the problem that the gradient disappears with an increase in the depth of the network. The feature vector outputted by the FCFE module is added to itself and inputted into the encoder for encoding after passing through the CSP feature fusion module.

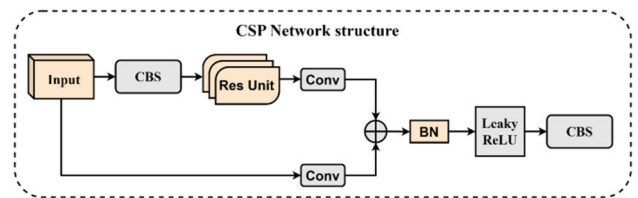


FIGURE 10. Diagram CSP network structure.

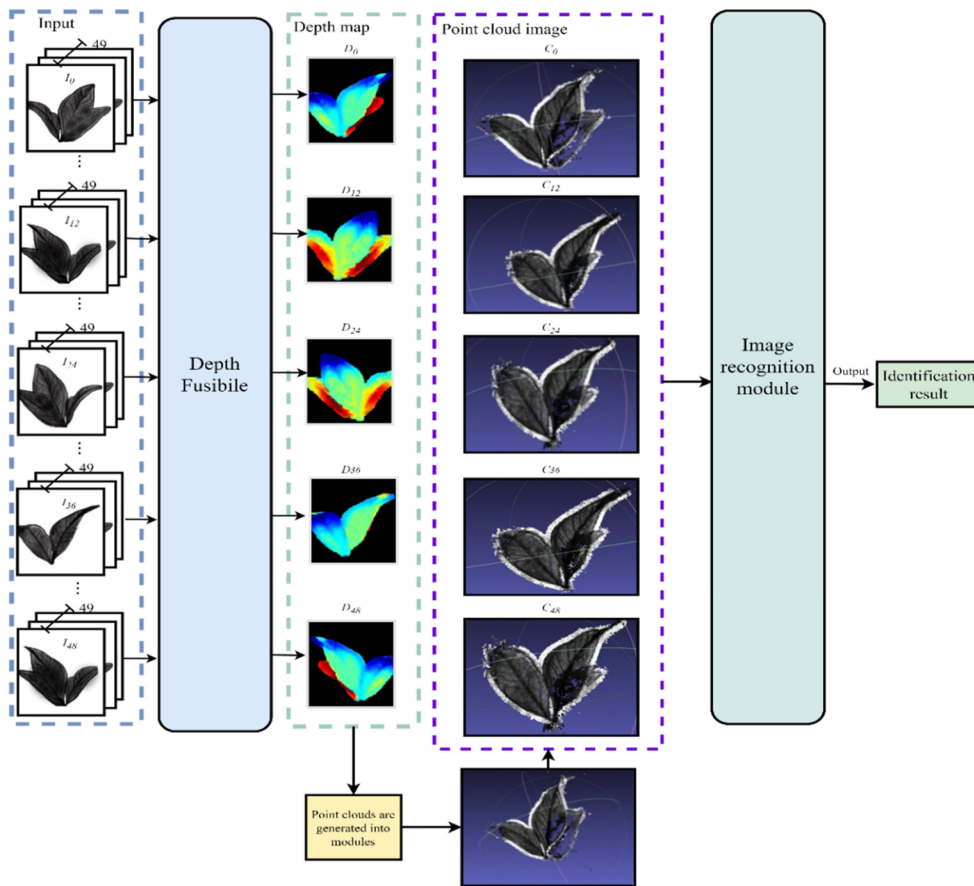
The algorithm for point-cloud-structured image object detection directly predicts its final detection results through the prediction module, outputting the category and location information of all targets in the current image. The prediction module is composed of a fully connected neural network, which is mainly divided into two branches: one for predicting the category of the target, and the other for predicting the location of the target.

#### IV. EXPERIMENTAL RESULTS AND ANALYSIS

##### A. TRAINING PROCESS

The training process of the algorithm for 3D THz transmission image reconstruction and object detection can be decomposed into two interrelated tasks: 3D reconstruction and object detection.

The network training set contained the terahertz dataset, the X-ray dataset and the DTU dataset, with each dataset containing images of different scenes. As shown in Figure 11, each main view was calculated with images from other perspectives to match the scores, and the top 49 images with high scores were obtained as the input of the network. Meanwhile, the size of the original images was reduced to  $400 \times 300$ , and images with a  $320 \times 256$  resolution were obtained by cropping from the center and used as the input dataset for training, while the camera's parameters were changed accordingly. The depth intervals refer to areas in the scene where there were changes in depth or discontinuities, which play an important role in 3D reconstruction and are crucial for obtaining accurate depth information and generating realistic 3D models. We set the assumed depth value to 425920.5, with a depth interval of 2.5 mm ( $D=192$  depth intervals).



**FIGURE 11.** Three-dimensional reconstruction of terahertz transmission images and training process of the target detection algorithm.

The feature of the main view and 48 other views were extracted by the transformer to obtain 49 feature maps. The feature maps and the camera's parameters were transformed by the homography matrix to form 49 feature cost bodies  $V_i$ . The variance of the points on the same space position of the 49 feature cost bodies was calculated, and  $V_i$  was aggregated into a cost space  $C$ . The dimensions of the cost space  $C$  were  $D$ ,  $W$ ,  $H$  and  $F$ , which are the number of depth samples, the width and height of the input dataset and the number of feature channels, respectively.

The value of each depth in the cost space was concentrated into a unimodal distribution. Each point was normalized using softmax along the direction of depth  $D$  to obtain the depth probability value of each point on  $D$ , thus obtaining a probability space  $P$ . Then the probability sum was calculated once for each four neighborhoods along the depth dimension of the probability space, and then the maximum probability sum was obtained along the depth dimension of  $D$ . The depth value of each point was calculated to obtain the depth map. The point cloud information was obtained after optimization of the fusion of 49 depth maps; the fusion process did not require training.

The task of detecting objects in terahertz images is to use a single neural network to act on the image, divide the image

into two regions and predict the probability of the boundary box and each region.

### B. EVALUATION OF THE ALGORITHM

Firstly, the 3D reconstruction part was trained for 16 epochs in total. The loss function of the training process is shown in Figure 12.

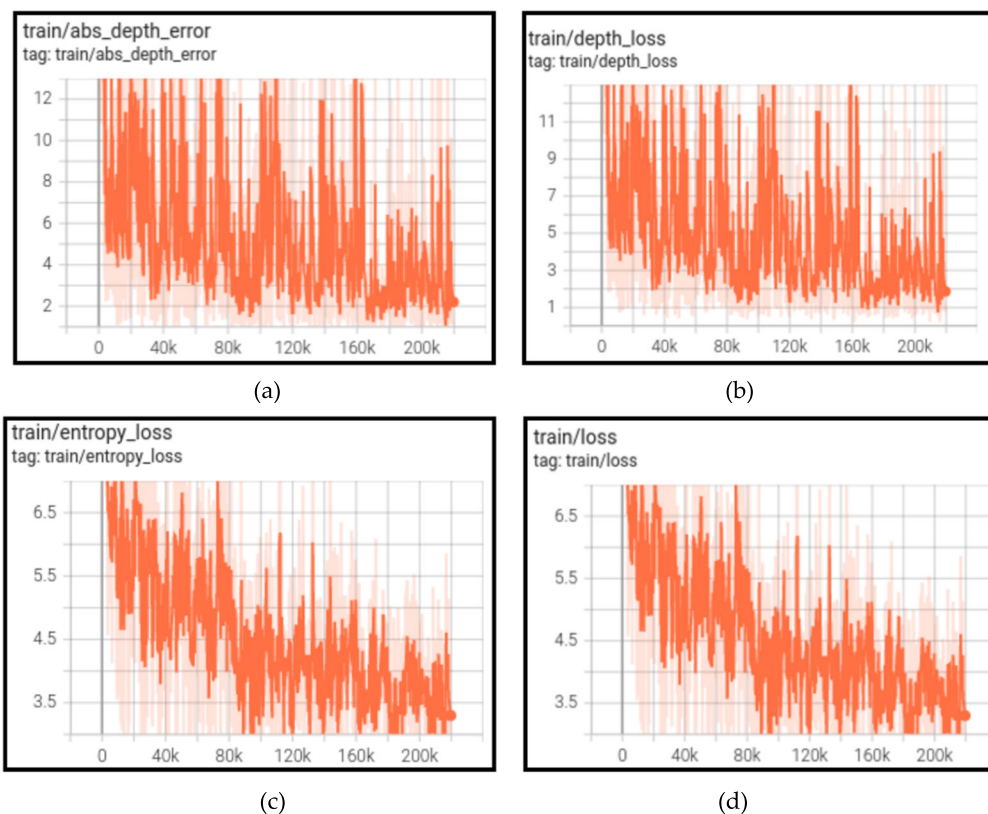
As can be seen from the figure, the loss function showed a general downward trend, indicating that the model had good learning ability. After completing the training of the 3D reconstruction part, the weights of the network were directly read to test the THz transmission separation images. The generated depth map is shown in Figure 13.

The depth maps generated from the multi-view terahertz images were fused and optimized to form a point cloud, and the results are shown in Figure 14:

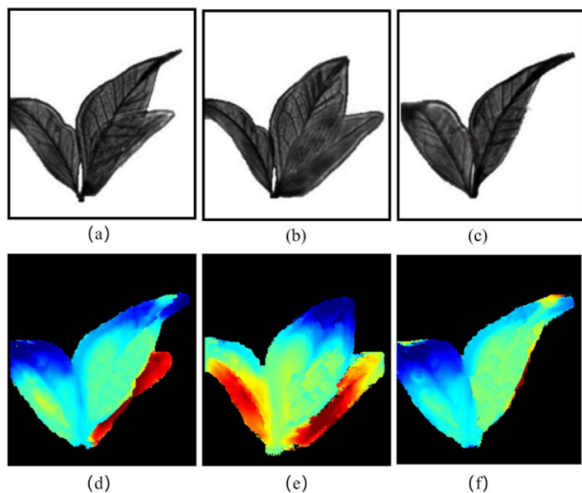
After the training of the 3D reconstruction of the terahertz images was complete, the detection network of the point-cloud-structured terahertz image of the target was trained. After the training was complete, the weights were directly read to test the point clouds of the terahertz images.

The test results are shown in Figure 15. It can be seen from the figure that the network could recognize and detect the structured point cloud of the terahertz images.





**FIGURE 12.** Change curves during training of the loss function of the algorithm for 3D reconstruction using terahertz transmission overlap separation: (a) declining curve of the loss in the depth map's error; (b) declining curve of the loss of the depth map loss; (c) descending curve of cross-entropy loss; (d) descending curve of total loss.



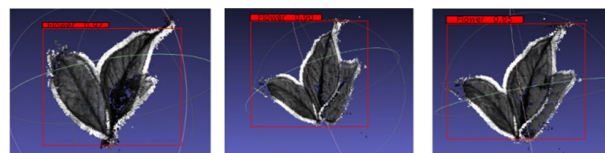
**FIGURE 13.** Results of the test of the algorithm for 3D terahertz transmission reconstruction : (a-f) Depth maps for Viewpoints 1-3.

To validate the superior reconstruction capacity of the FCTMVSNet algorithm, this paper compared the reconstruction accuracy of various algorithms based on the DTU dataset, as demonstrated in Table 1.

Based on the leaf terahertz image dataset collected through the experiment, the FCTMVSNet 3D reconstruction



**FIGURE 14.** Point cloud generated by fusion of the depth map.

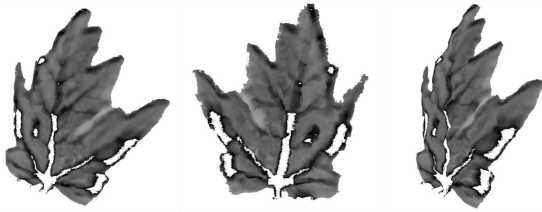
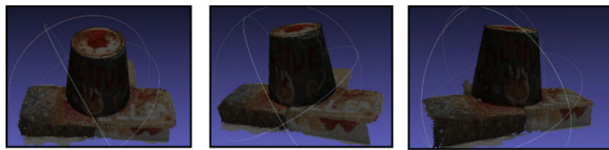
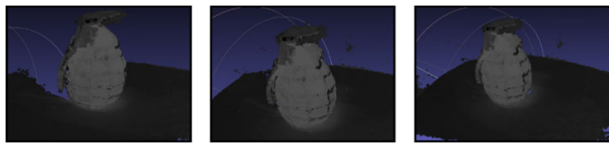


**FIGURE 15.** Results of detection and recognition of the point cloud map of the terahertz images.

algorithm proposed in this paper is verified, and the reconstruction effect is presented in Figure 15. It can be observed from the reconstruction effect diagram that the FCTMVSNet reconstruction algorithm proposed in this paper enables 3D reconstruction on the terahertz dataset, and the reconstructed leaves possess a relatively complete structure. However, due to the low quality of the images collected in the experiment, part of the internal texture of the reconstructed leaves is missing. Subsequently, by adjusting the experimental parameters

**TABLE 1. Comparative analysis of the reconstruction accuracy of different algorithms based on the DTU dataset.**

Method	Acc.(mm)	Comp.(mm)	Overall.(mm)
COLMAP	0.412	0.667	0.539
TransMVSNet	0.331	0.292	0.312
GC-MVSNet	0.352	0.284	0.318
FCTMVSNet	0.321	0.294	0.308

**FIGURE 16. Reconstruction of the Terahertz dataset.****FIGURE 17. Reconstruction of the DTU dataset.****FIGURE 18. Reconstruction of the X-ray dataset.**

of terahertz imaging, superior terahertz image data can be obtained, and the 3D reconstruction effect can be optimized.

At the same time, the FCTMVSNet algorithm proposed in this paper is also reconstructed on the DTU public dataset and the X-ray dataset, and can demonstrate an excellent reconstruction effect on both datasets, as depicted in Figures 17 and 18.

It can be seen from the figure that whether we used the DTU dataset with the camera's parameters, or the X-ray and terahertz datasets with the camera's parameters calculated by the shooting angles, FCTMVSNet could reconstruct objects from multi-angle images and achieve good results.

## V. CONCLUSION

In this study, we proposed a Transformer-based 3D reconstruction algorithm known as FCTMVSNet (feature and cost transformer depth inference for unstructured multi-view stereo). We used a self-attention mechanism to replace the traditional convolution-based feature extraction network, which solved the problem that convolution is limited by the information on the spatial location and cannot obtain the feature information between images. At the same time, we proposed an inter-layer attention mechanism of the cost body to extract the context information between cost layers and improve the model's accuracy. When verified on public databases,

the FCTMVSNet network successfully performed 3D reconstruction of the processed terahertz images. On the basis of obtaining the point cloud of the measured object through terahertz images, we structured the pose angle and parameters to output the fixed direction of the point cloud of the texture for target detection. The test results showed that the generated structured terahertz images met the requirements of target detection, providing theoretical support for subsequent 3D terahertz imaging.

## ACKNOWLEDGMENT

Lun Li and Xiaojin Wu: conception and design of the work, acquisition, analysis, and interpretation of the data. Xudong Lu: code programming and tests. Fan Bai: code programming and testing, design of the optimization scheme. Yuan Gao: collected the references. Wencheng Wang: drafted the work and revised it critically for important intellectual content. Haixian Liu: polished the manuscript including language and grammar.

## REFERENCES

- [1] F. Bai, L. Li, W. Wang, and X. Wu, "DETransMVSNet: Research on terahertz 3D reconstruction of multi-view stereo network with deep equilibrium transformers," *IEEE Access*, vol. 11, pp. 146042–146053, 2023, doi: [10.1109/ACCESS.2023.3342847](https://doi.org/10.1109/ACCESS.2023.3342847).
- [2] H. Cheon, H.-J. Yang, and J.-H. Son, "Toward clinical cancer imaging using terahertz spectroscopy," *IEEE J. Sel. Topics Quantum Electron.*, vol. 23, no. 4, pp. 1–9, Jul. 2017.
- [3] A. Gong, Y. Qiu, X. Chen, Z. Zhao, L. Xia, and Y. Shao, "Biomedical applications of terahertz technology," *Appl. Spectrosc. Rev.*, vol. 55, no. 5, pp. 418–438, 2019.
- [4] W. Wang, X. Yuan, X. Wu, and Y. Liu, "Fast image dehazing method based on linear transformation," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1142–1155, Jun. 2017, doi: [10.1109/TMM.2017.2652069](https://doi.org/10.1109/TMM.2017.2652069).
- [5] Z. Zhang and T. Buma, "Adaptive image reconstruction for sparse arrays using single-cycle terahertz pulses," *Opt. Lett.*, vol. 35, no. 10, p. 1680, 2010.
- [6] E. Abraham, A. Younus, C. Aguerre, P. Desbarats, and P. Mounaix, "Refraction losses in terahertz computed tomography," *Opt. Commun.*, vol. 283, no. 10, pp. 2050–2055, May 2010.
- [7] D. Robertson, P. Marsh, and D. Bolton, "340-GHz 3D radar imaging test bed with 10 Hz frame rate," *Proc. SPIE*, vol. 8362, pp. 41–51, May 2012, doi: [10.1117/12.918581](https://doi.org/10.1117/12.918581).
- [8] H. Li, C. Li, S. Wu, S. Zheng, and G. Fang, "Adaptive 3D imaging for moving targets based on a SIMO InSAR imaging system in 0.2 THz band," *Remote Sens.*, vol. 13, no. 4, p. 782, Feb. 2021, doi: [10.3390/rs13040782](https://doi.org/10.3390/rs13040782).
- [9] S. R. Tripathi, Y. Sugiyama, K. Murate, K. Imayama, and K. Kawase, "Terahertz wave three-dimensional computed tomography based on injection-seeded terahertz wave parametric emitter and detector," *Opt. Exp.*, vol. 24, no. 6, p. 6433, Mar. 2016.
- [10] T. Zhou, R. Zhang, C. Yao, Z.-L. Fu, D.-X. Shao, and J.-C. Cao, "Terahertz three-dimensional imaging based on computed tomography with photonics-based noise source," *Chin. Phys. Lett.*, vol. 34, no. 8, Aug. 2017, Art. no. 084206.
- [11] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Comput. Vis. (ECCV)*, 2018, pp. 785–801, doi: [10.1007/978-3-030-01237-3\\_47](https://doi.org/10.1007/978-3-030-01237-3_47).
- [12] K. Zhang, M. Liu, J. Zhang, and Z. Dong, "PA-MVSNet: Sparse-to-dense multi-view stereo with pyramid attention," *IEEE Access*, vol. 9, pp. 27908–27915, 2021, doi: [10.1109/ACCESS.2021.3058522](https://doi.org/10.1109/ACCESS.2021.3058522).
- [13] W. Wang, Z. Chen, X. Yuan, and X. Wu, "Adaptive image enhancement method for correcting low-illumination images," *Inf. Sci.*, vol. 496, pp. 25–41, Sep. 2019, doi: [10.1016/j.ins.2019.05.015](https://doi.org/10.1016/j.ins.2019.05.015).
- [14] D. McGonigle, T. Wang, J. Yuan, K. He, and B. Li, "I2S2: Image-to-scene sketch translation using conditional input and adversarial networks," in *Proc. IEEE 32nd Int. Conf. Tools With Artif. Intell. (ICTAI)*, Nov. 2020, pp. 773–778, doi: [10.1109/ICTAI50040.2020.00123](https://doi.org/10.1109/ICTAI50040.2020.00123).

- [15] W. Wang, D. Yan, X. Wu, W. He, Z. Chen, X. Yuan, and L. Li, "Low-light image enhancement based on virtual exposure," *Signal Process., Image Commun.*, vol. 118, Oct. 2023, Art. no. 117016, doi: 10.1016/j.image.2023.117016.
- [16] W. Wang, X. Wu, X. Yuan, and Z. Gao, "An experiment-based review of low-light image enhancement methods," *IEEE Access*, vol. 8, pp. 87884–87917, 2020, doi: 10.1109/ACCESS.2020.2992749.
- [17] J.-B. Perraud, A. Chopard, J.-P. Guillet, P. Gellie, A. Vuillot, and P. Mounaix, "A versatile illumination system for real-time terahertz imaging," *Sensors*, vol. 20, no. 14, p. 3993, Jul. 2020, doi: 10.3390/s20143993.
- [18] W. J. Liu, J. K. Wang, and H. C. Qu, "Multi-scale cost volumes information sharing based multi-view stereo reconstructed model," *J. Image Graph.*, vol. 27, no. 11, pp. 3331–3342, 2022.
- [19] Y. Li, Z. Zhao, J. Fan, and W. Li, "ADR-MVSNet: A cascade network for 3D point cloud reconstruction with pixel occlusion," *Pattern Recognit.*, vol. 125, May 2022, Art. no. 108516.
- [20] Y. Xu and U. Stilla, "Toward building and civil infrastructure reconstruction from point clouds: A review on data and key techniques," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2857–2885, 2021.
- [21] Z. Yu and S. Gao, "Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and Gauss–Newton refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1946–1955.
- [22] H. J. Liu, "Fusion attention mechanism and multilayer U-Net for multi-view stereo," *J. Image Graph.*, vol. 27, no. 2, pp. 475–485, 2022.
- [23] Z. Chuang, W. Wei, and D. Yuxuan, "Neural collaborative recommendation algorithm based on attention mechanism and knowledge graph," *Comput. Eng. Appl.*, vol. 27, no. 2, pp. 475–485, 2022.
- [24] J. Yimu, T. Shuning, and L. Shangdong, "A dual attention mechanism based on TransH for distantly-supervised relation extraction," *J. Nanjing Univ. Posts Telecommun. Natural Sci. Ed.*, vol. 6, pp. 70–78, Jan. 2022(06):70-78.

**HAIXIAN LIU**, photograph and biography not available at the time of publication.

**FAN BAI**, photograph and biography not available at the time of publication.

**XUDONG LU**, photograph and biography not available at the time of publication.

**YUAN GAO**, photograph and biography not available at the time of publication.



**XIAOJIN WU** received the Ph.D. degree in traffic information engineering control from Beijing Jiaotong University, in 2012. He is currently with the School of Machinery and Automation, Weifang University. He is also a Visiting Scholar with the University of North Texas, engaged in the research of image enhancement technology. He has published and authored more than 20 articles on academic journals and conferences. His main research interests include intelligent systems and image processing.



**LUN LI** was born in 1989. He received the Doctor of Engineering degree in weapon science and technology from Shenyang Polytechnic University, in July 2020. He is currently a Lecturer with the School of Information and Control Engineering, Weifang University, mainly engaged in the research of terahertz imaging detection technology and bionic robot control.

...