



Making the Invisible Visible: Toward High-Quality Terahertz Tomographic Imaging via Physics-Guided Restoration

Weng-Tai Su¹ · Yi-Chun Hung¹ · Po-Jen Yu¹ · Shang-Hua Yang¹ · Chia-Wen Lin¹

Received: 1 February 2022 / Accepted: 26 April 2023 / Published online: 7 June 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Terahertz (THz) tomographic imaging has recently attracted significant attention thanks to its non-invasive, non-destructive, non-ionizing, material-classification, and ultra-fast nature for object exploration and inspection. However, its strong water absorption nature and low noise tolerance lead to undesired blurs and distortions of reconstructed THz images. The diffraction-limited THz signals highly constrain the performances of existing restoration methods. To address the problem, we propose a novel multi-view Subspace-Attention-guided Restoration Network (SARNet) that fuses multi-view and multi-spectral features of THz images for effective image restoration and 3D tomographic reconstruction. To this end, SARNet uses multi-scale branches to extract intra-view spatio-spectral amplitude and phase features and fuse them via shared subspace projection and self-attention guidance. We then perform inter-view fusion to further improve the restoration of individual views by leveraging the redundancies between neighboring views. Here, we experimentally construct a THz time-domain spectroscopy (THz-TDS) system covering a broad frequency range from 0.1 to 4 THz for building up a temporal/spectral/spatial/material THz database of hidden 3D objects. Complementary to a quantitative evaluation, we demonstrate the effectiveness of our SARNet model on 3D THz tomographic reconstruction applications.

Keywords Terahertz imaging · Image restoration · Computed tomography · Deep learning · Self-attention

1 Introduction

Ever since the first camera's invention, imaging under different bands of electromagnetic (EM) waves, especially X-ray and visible lights, has revolutionized our daily lives (Kamruzzaman et al., 2011; Rotermund et al., 1991; Yujiri et al., 2003). X-ray imaging plays a crucial role in medical diagnoses, such as cancer, odontopathy, and COVID-19 symptom (Abbas et al., 2021; Round et al., 2005; Tuan et al., 2018), based on

X-ray's high penetration depth to great varieties of materials; visible-light imaging has not only changed the way of recording lives but contributes to the development of artificial intelligence (AI) applications, such as surveillance security and surface defect inspection (Xie, 2008). However, X-ray and visible-light imaging still face tough challenges. X-ray imaging is ionizing, which is harmful to biological objects and thus severely limits its application scope (de Gonzalez and Darby, 2004). On the other hand, although both non-ionizing and non-destructive, visible-light imaging cannot retrieve interior information of most objects which are opaque in visible light due to the highly absorptive and intense scattering behaviors between light and matter in the visible light band. To visualize the 3D information of objects in a remote but accurate manner, terahertz (THz) imaging has become among the most promising candidates among all EM wave-based imaging techniques (Abraham et al., 2010; Calvin et al., 2012).

Table 1 shows the comparison of different types of high-resolution imaging modalities with a non-contact setting. As camera and Light Detection and Ranging (LiDAR) are widely launched for 2D/3D image capturing, due to the

✉ Chia-Wen Lin
cwlin@ee.nthu.edu.tw

Weng-Tai Su
wengtai2008@hotmail.com

Yi-Chun Hung
nick831111@gmail.com

Po-Jen Yu
jerry321ab@gmail.com

Shang-Hua Yang
shanghua@ee.nthu.edu.tw

¹ Department of Electrical Engineering, National Tsing Hua University, Kuang-Fu Road, Hsinchu 30048, Taiwan

Table 1 Comparison of features of existing imaging technologies. The ability to see through objects opaque in visible light enables tomography. The X-ray would ionize objects, which means not bio-safe. Some methods can identify different materials by their spectroscopy, and they

require the penetration of the object. One imaging method is more favorable if it can be placed on the table (table-top), thereby excluding those methods which require bulky instruments and placed in a special room such as X-ray and Magnetic Resonance Imaging (MRI)

Method	See through opaque objects	Bio-safe	Material identification	Table-top system
RGB Camera	×	✓	Partially ^a	✓
X-ray	✓	×	✓	✓
LiDAR	×	✓	Partially ^b	✓
Ultrasonic	✓	✓	✓	✓
MRI	✓	✓	Partially ^c	×
THz Imaging	✓	✓	✓	✓

^aMaterial of object surface (fabric, plastic, wood, paper, leather, metal, and fur) (Kim et al., 2018)

^bMaterial of object surface (fabric, brick, pine, wood, and maple leaves) (Nunes-Pereira et al., 2020)

^cMaterial with hydrogen atoms (tumor, fat, and water) (Clarke et al., 1995)

intensive scattering and absorption happening nearby object surfaces, these two imaging methods cannot visualize 3-D full profiles of most objects. Research groups have successfully found other electromagnetic spectrum regimes to bring information invisible to visible to address this issue. X-ray imaging is one of the commonly used methods to precisely visualize the interior of objects (Chapman et al., 1997; Fitzgerald, 2000; Sakdinawat, A., Attwood, D., 2010; Cloetens et al., 1996; Peterson et al., 2001). Despite its invisible-to-visible capability, high-energy X-ray photons would cause both destructive and ionizing impacts on various material types preventing further investigations with other material characterization modalities. Magnetic resonance imaging (MRI) technology has proven to be a bio-safe way to visualize soft materials with excellent image contrast. Still, it is bulky and requires sufficient space for operation, which prevents its practical use in many application scenarios. To be pervasively used like visible light cameras, the desired tomographic imaging modality must be operated at a remote distance, non-destructive, bio-safe, compact, and most importantly, capable of digging out information conventional cameras cannot achieve.

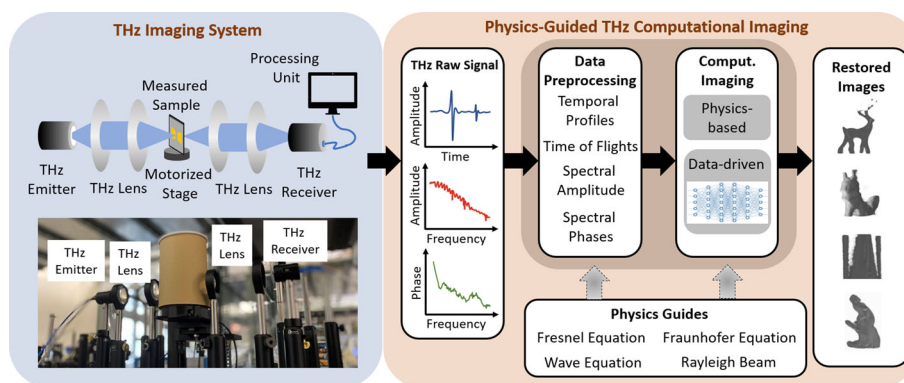
THz radiation, between microwave and infrared, has often been regarded as the last frontier of EM wave (Saeedkia, 2013), which provides its unique functionalities among all EM bands. Along with the rapid development of THz technology, THz imaging has recently attracted significant attention due to its non-invasive, non-destructive, non-ionizing, material-classification, and ultra-fast nature for advanced material exploration and engineering. As THz waves can partially penetrate through varieties of materials being opaque in visible light, they carry hidden material tomographic information along the traveling path, making this approach a desired way to see through black boxes without damaging the exterior (Mittleman et al., 1999; Jansen et al., 2010; Mittleman, 2018). By utilizing light-matter interaction within the THz band, multifunctional tomographic information from

a great variety of materials can also be retrieved even at a remote distance. In the past decades, THz time-domain spectroscopy (THz-TDS) has become one of the most representative THz imaging modalities to achieve non-invasive evaluation because of its unique capability of extracting geometric and multi-functional information of objects. Owing to its fruitful information in multi-dimensional domains—space, time, frequency, and phase, THz-TDS imaging has been already allocated for numerous emerging fields, including drug detection (Kawase et al., 2003), industrial inspection, cultural heritage inspection (Fukunaga, 2016), and cancer detection (Bowman et al., 2018).

However, the conventional methods (e.g., Time-max (Hung & Yang, 2019b)) for THz imaging is to analyze the temporal profiles of THz signals measured by a THz-TDS system within a limited time window. The reconstructed tomographic image quality is severely constrained by the diffraction-limited geometry and absorption behavior of objects in the THz spectral regime, leading to undesired blurring and distortion of reconstructed tomography images. To address this problem, we utilize useful spectral bands to supplement the conventional method, recording the maximum amplitude of the time-domain THz signal of each pixel for recovery of the clear 2D images.

Recently, data-driven methods based on deep learning models, which do not resort to any explicit transform model but are learned from representative big data, have been revolutionizing the physics-based paradigm in image restoration. The data-driven methods can be regardless of physical properties while maintaining the advantages of the physics-based methods and achieving state-of-the-art performances. We can also cast THz image analysis as an image-domain learning problem. Nevertheless, a THz image retrieved from THz raw time-domain signals does not carry enough restoration information, thereby limiting the efficacy of the data-driven methods. Furthermore, we found that directly learning from the **full spectral information** to restore THz images leads

Fig. 1 Flowchart of physics-guided THz computational imaging. The pixel-wise THz raw signals are measured from the THz imaging system along with image data. The multi-domain data are then processed and fused by a computational imaging model to reconstruct the images. The computational imaging model can be either physics-based or data-driven, with or without the physics guides derived from the physical properties of THz signals



to unsatisfactory performance. The main reason is that the full spectra of THz signals involve diverse characteristics of materials, noises, and scattered signals, which causes difficulties in model training. To address the above issues, as illustrated in Fig. 1, we can leverage additional pixel-wise spectral information carried in the THz raw signals, such as the amplitude/phase spectra corresponding to specific physical characteristics of THz waves passing through materials. Due to a large number of spectral bands with measured THz image data, it is desirable to sample a subset of the most physics-prominent spectral bands to reduce the number of training parameters. Specifically, The THz beam is significantly attenuated at water absorption frequencies. As a result, such physics-guided water-absorption property of THz beams offers useful clues for inspecting and reconstructing an object from THz images captured in a see-through setting (e.g., computed tomographic reconstruction) as will be elaborated in Sect. 5.2.

Based on the concept revealed in Fig. 1, we here propose a multi-scale **Subspace-Attention** guided **Restoration Network** (SARNet) that fuses intra-view complementary spectral features of the THz amplitude and phase to supplement the *Time-max* image for restoring clear 2D images. To this end, SARNet learns common representations in a common latent subspace shared between the amplitude and phase, and then incorporates a Self-Attention mechanism to learn the wide-range dependency of the spectral features for guiding the restoration task. To leverage the inter-view redundancies existing between neighboring views of an object captured from different angles, on top of SARNet we also propose a multi-view version image restoration model, namely SARNet_{MV}, that incorporates inter-view fusion to further boost restoration performance. Finally, from clear 2D views restored from the corrupted views of an object, we can reconstruct high-quality 3D tomography via inverse Radon transform. Our main contributions are summarized as follows:

- We are the first research group to merge THz temporal-spatial-spectral data, data-driven models, and light-

matter interaction properties to the best of our knowledge. The proposed SARNet achieves excellent performance in extracting and fusing features from the light-matter interaction data in THz spectral regime, which inherently contains fruitful 3D object information and its material behaviors. Based on the architecture of the proposed SARNet_{MV} on intra/inter-view feature fusion, it delivers state-of-the-art performance on THz image restoration.

- With our newly established THz-TDS tomography dataset—the world’s first in its kind, we provide comprehensive quantitative/qualitative analyses among SARNet_{MV} and state-of-the-arts. SARNet_{MV} significantly outperforms *Time-max* (Hung & Yang, 2019b), U-Net (Ronnerberger et al., 2015), and NENet (Cheng et al., 2021) by 11.41 dB, 2.79 dB, and 2.23 dB, respectively, in average PSNR at reasonable computation and memory costs.
- This work shows that computer vision techniques can significantly contribute to the THz community and further open up a new interdisciplinary research field to boost practical applications, e.g., non-invasive evaluation, gas tomography, industrial inspection, material exploration, and biomedical imaging.

2 Related Work

2.1 Conventional THz Computational Imaging

In the past decades, many imaging methods have been developed based on the light-matter interaction in the THz frequency range. Based on THz absorption imaging modalities, the material refractive index mapping can be profiled through Fresnel equation (Born & Wolf, 2013), extracted by the THz power loss while propagating through the tested object boundary. With THz spectroscopy imaging, both material information encoded in the wave propagation equation and object geometry can be revealed. To be more specific, the depth map of the measured object can be reconstructed based on the phase spectrum of the retrieved THz signals (Hack and Zolliker, 2014); the attenuated power spectrum

information can further recover the hyperspectral material fingerprint mapping. These characteristics provide functional 3D imaging capability for object inspection. Additionally, considering the propagated THz beam behavior of a signal as the model prior knowledge, such as Rayleigh beam, has proven to largely improve the imaging quality (Recur et al., 2012). With the THz time-reversal techniques, the THz amplitude and/or phase images of a measured object can be estimated by the spatiotemporal interaction between the input THz waves and the object. However, the application scopes of those physics-driven methods are severely limited since they normally require a sufficient amount of prior knowledge of a measured object to simplify the guided complex physical models. To break this limitation, data-driven approaches, especially deep neural networks, start to arouse intensive attention due to their excellent learning capability. A data-driven model based on physical priors can effectively loosen the requirement of prior knowledge of materials and perform superior to conventional physics-based methods. Moreover, data-driven models can learn to adequately fuse the different information of THz signals, such as amplitude/phase spectra and the time-resolved THz signals, to achieve superior image restoration (Su et al., 2022, 2023).

2.2 Physics-Guided Data-Driven THz Imaging

In contrast to those model-based methods, data-driven methods are mainly based on deep learning models (Zhang et al., 2017; Mao et al., 2016), which do not resort to any explicit transform model but are learned from representative big data. We can cast THz image analysis as an image-domain learning problem. Deep learning has revolutionized the aforementioned physics-based paradigm in image restoration, for which the data-driven methods can be regardless of physical properties while maintaining the advantages of the physics-based methods and achieving state-of-the-art performances. Nevertheless, a THz image retrieved from THz raw time-domain signals does not carry enough restoration information, thereby limiting the efficacy of the data-driven methods. To address the issue, as illustrated in Fig. 1, we can leverage additional pixel-wise spectral information carried in the THz raw signals, such as the amplitude/phase spectra corresponding to specific physical characteristics of THz waves passing through materials. By contrast, the physics-based methods are difficult to leverage such pixel-wise amplitude/phase spectral information. To this end, the data-driven model proposed in Su et al. (2022) incorporates additional information from amplitude/phase at water absorption frequencies, derived from the physical properties of THz signals, to complement the insufficient information in time-domain THz images so as to significantly boost restoration performance. In addition, if the THz imaging system uses the THz focal beam, the THz beam diameter along with the

wave propagation direction can be varied. Additionally, the THz beam diameter can also be changed in different spectral bands due to the diffraction limit. Both changed THz beam diameters lead to the non-identical point spread function (PSF) in each measurement point. To solve this problem, the Filter Adaptive Convolutional Layer (FAC) (Zhou et al., 2019) can learn different filter kernels corresponding to the PSF for each pixel from spatial-spectral information and use those kernels to deliver superior imaging performance.

2.3 Deep Learning-Based Image Restoration

In recent years, deep learning methods were first popularized in high-level visual tasks, and then gradually penetrated into many tasks such as image restoration and segmentation. Convolutional neural networks (CNNs) have proven to achieve state-of-the-art performances in fundamental image restoration problems (Mao et al., 2016; Zhang et al., 2017, 2020, 2018; Ronneberger et al., 2015). Several network models for image restoration were proposed, such as U-Net (Ronneberger et al., 2015), hierarchical residual network (Mao et al., 2016) and residual dense network (Zhang et al., 2020). Notably, DnCNN (Zhang et al., 2017) uses convolutions, BN, and ReLU to build 17-layer network for image restoration which was not only utilized for blind image denoising, but was also employed for image super-resolution and JPEG image deblocking. FFDNet (Zhang et al., 2018) employs noise level maps as inputs and utilizes a single model to develop variants for solving problems with multiple noise levels. In Mao et al. (2016) a very deep residual encoding-decoding (RED) architecture was proposed to solve the image restoration problem using skip connections. (Zhang et al., 2020) proposed a residual dense network (RDN), which maximizes the reusability of features by using residual learning and dense connections. NBNNet (Cheng et al., 2021) employs subspace projection to transform learnable feature maps into the projection basis, and leverages non-local image information to restore local image details. Similarly, the Time-max image obtained from a THz imaging system can be cast as an image-domain learning problem which was rarely studied due to the difficulties in THz image data collection. Research works on image-based THz imaging include (Popescu and Ellicar, 2010; Popescu et al., 2009; Wong et al., 2019), and THz tomographic imaging works include (Hung & Yang, 2019b, a).

Transformer (Vaswani et al., 2017), a kind of self-attention mechanism for machine learning, was first proposed to largely boost the research in natural language processing. Recently, it has gained wide popularity in the computer vision community, such as image classification (Dosovitskiy et al., 2020; Wu et al., 2020), object detection (Carion et al., 2020; Liu et al., 2018), segmentation (Wu et al., 2020), which learns to focus on essential image regions by exploring the long-

range dependencies among different regions. Transformer has also been introduced for image restoration (Chen et al., 2021; Cao et al., 2021; Wang et al., 2022) due to its impressive performance. In Chen et al. (2021), a standard Transformer-based backbone model IPT was proposed to address various restoration problems, which relies on a large number of parameters (over 115.5M parameters), large-scale training datasets, and multi-task learning for achieving high restoration performances. Additionally, VSR-Transformer (Cao et al., 2021) first utilizes a CNN to extract visual features and then adopts a self-attention model to fuse features for video super-resolution. Although transformer-based attention mechanisms have proven effective in boosting the performance of image restoration tasks, the performance gains of transformers come at the cost of significantly larger amounts of training data and computation.

2.4 Tomographic Reconstruction

Computed tomographic (CT) imaging methods started from X-ray imaging, and many methods of THz imaging are similar to those of X-ray imaging. One of the first works to treat X-ray CT as an image-domain learning problem was Kang et al. (2017), which adopts CNN to refine tomographic images. In Jin et al. (2017), U-Net was used to refine image restoration with significantly improved performances. Zhu et al. (2018) further projects sinograms measured directly from X-ray into higher-dimensional space and uses domain transfer to reconstruct images. The aforementioned works were specially designed for X-ray imaging.

Hyperspectral imaging (Schultz et al., 2001; Ozdemir and Polat, 2020; Geladi et al., 2004) constitutes image modalities other than THz imaging. Different from THz imaging, Hyperspectral imaging collects continuous spectral band information of the target sample. Typically, the frequency bands fall in the visible and infrared spectrum; hence, most hyperspectral imaging modalities can only observe the surface characteristics of targeted objects. Furthermore, although existing deep-based hyperspectral imaging works can learn spatio-spectral information from a considerable amount of spectral cube data, they mainly rely on the full spectral information to restore hyperspectral images. This would usually lead to unsatisfactory performance for THz imaging since the full spectral bands of THz signals involve diverse characteristics of materials, noises, and scattered signals, which causes difficulties in model training.

3 Physics-Guided THz Imaging

Based on the dependency between the amplitude of a temporal signal and THz electric field, in conventional THz imaging, the maximum peak of the signal (Time-max (Hung

&Yang, 2019b)) is extracted as the feature for a voxel. The reconstructed image based on Time-max features can deliver a great signal-to-noise ratio and a clear object contour. However, the conventional THz imaging based on Time-max features suffers from several drawbacks, such as the undesired contour in the boundary region, the hollow in the body region, and the blurs in high spatial-frequency regions. To break this limitation, we utilize the spectral information of THz temporal signals to supplement the conventional method based on Time-max features since the voxel of the material behaviors is encoded in both the phase and amplitude of different frequency components, according to the Fresnel equation (Dorney et al., 2001).

More specifically, considering an incident THz wave penetrates through a single-material object with thickness d , the detected THz signal $S_d(f)$ at frequency f is determined by the material complex refractive index $\tilde{n}_o(f) = n_o(f) - j\kappa_o(f)$ and the thickness d in (1).

$$\begin{aligned} S_d(f) &= S_{\text{ref}}(f) \cdot t(\tilde{n}_o, f) \cdot \exp\left[\frac{\kappa_o(f)2\pi fd}{c}\right] \cdot \\ &\quad \exp\left[\frac{-jn_o(f)2\pi fd}{c}\right] \\ &= S_{\text{ref}}(f) \cdot t(\tilde{n}_o, f) \cdot (I_a(f))^d \cdot (I_p(f))^d, \end{aligned} \quad (1)$$

where $S_{\text{ref}}(f)$ and $t(\tilde{n}_o, f)$ are respectively the THz input signal and the Fresnel loss of THz waves (e.g., amplitude attenuation and phase change) due to the air-object interface at frequency f . Here, the Fresnel loss resulting from the presence of a single material can be further simplified as a constant. Meanwhile, $I_a(f) = \exp\left[\frac{\kappa_o(f)2\pi f}{c}\right]$ and $I_p(f) = \exp\left[\frac{-jn_o(f)2\pi f}{c}\right]$ can be acquired in a data-driven manner using information regarding the object thickness (i.e., ground-truth) and the detected THz signal. Specifically, although the complex refractive index is not provided explicitly, the network can still learn to map noisy input amplitude/phase images to their corresponding ground-truth images.

To provide a more detailed explanation of THz imaging, Fig. 2 shows the flowchart of estimating amplitude and phase information of $S_d(f)$ from the raw data directly measured by the THz-TDS system. This figure illustrates time-domain THz signals measured in air, the body, and the leg of a 3-D printed deer, respectively. While the THz beam passes through the object, the attenuated THz time-domain signal encodes the thickness and material information of the THz-illuminated region. By processing the peak amplitudes of THz signals (i.e., Time-max), the 3-D profile of the printed deer can be further reconstructed. Although this conventional way is well-fitted for visualizing 3-D objects, the inherent diffraction behavior and strong water absorption nature of THz wave induce various kinds of noise sources as well as

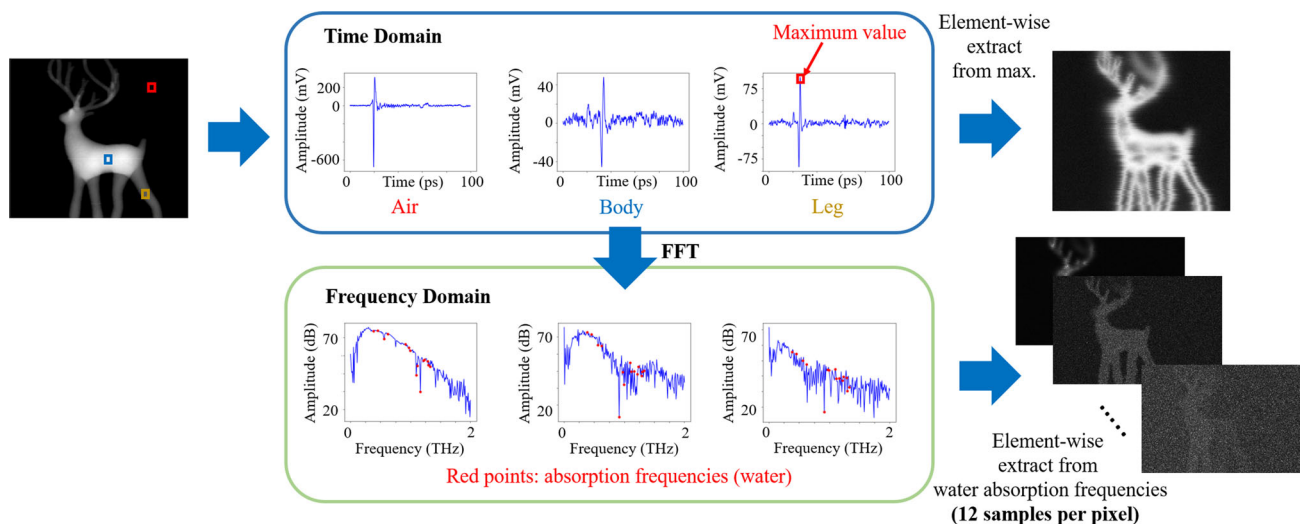


Fig. 2 Raw data of measured THz images. This figure illustrates the time domain data measured in air and the body and leg of our 3-D printed deer. The red points illustrate the frequency bands with strong water absorption. The right figures illustrate the reconstructed image using

the max value of the time domain (upper right), and the reconstructed image using different water absorption frequencies (lower right) (Color figure online)

the loss of material information, as characterized by parameters such as $t(\tilde{n}_o, f)$, $I_a(f)$, and $I_p(f)$ in (1). This leads to the undesirable blurring, distorted, speckled phenomenon of functional THz images. Existing works have tackled this issue to restore clear images via estimating point spread functions (Popescu and Ellicar, 2010; Popescu et al., 2009), image enhancement (Wong et al., 2019), machine learning (Ljubenovic et al., 2020; Wong et al., 2019), and more. Their performance is, however, still severely constrained by diffraction-limited THz beam. To break the limitations, the motivation of our work is to reconstruct deep-subwavelength tomographic images by using a deep-learning-based image restoration method and spatio-spectral information of the hidden objects.

3.1 Water Absorption Profile-Guided THz Imaging

As shown in Fig. 2, each 2-D THz image is composed of an array of time-domain signals, from which the Fourier transform operation can be utilized to extract voxel-wise multi-spectral features. Due to a large number of spectral bands with measured THz image data, it is required to sample a small subset of prominent spectral bands to reduce the training burden. Because the THz wave is significantly attenuated at water absorption frequencies, selecting THz bands based on water absorption lines can better delineate an object and characterize its thickness profile. The spectral information, including both amplitude and phase at the selected frequencies, is extracted and then employed to restore clear 2D images. The different features in THz images at THz water-absorption frequencies (the 12 selected frequencies in this

work: 0.380, 0.448, 0.557, 0.621, 0.916, 0.970, 0.988, 1.097, 1.113, 1.163, 1.208, and 1.229 THz) as shown in Fig. 3. It shows multiple 2D THz images of the same object at the selected frequencies, showing very different contrasts and spatial resolutions as these hyperspectral THz image sets have different physical characteristics through the interaction of THz waves with objects.

The lower-frequency phase images offer relatively accurate depth information due to their higher SNR level, whereas the higher-frequency phase images offer finer contours and edges because of the shrinking diffraction-limited wavelength sizes (from left to right in Fig. 3). The phase also contains, however, a great variety of information on light-matter interaction that could cause learning difficulty for the image restoration task. To address this issue, we utilize the amplitude spectrum as complementary information. Although the attenuated amplitude spectrum cannot reflect comparable depth accuracy levels as the phase spectrum, the amplitude spectrum still presents superior SNR and more faithful contours such as the location information of a measured object. Besides, as the complementary information to phase, the lower-frequency amplitude offers higher contrast, whereas the higher-frequency amplitude offers a better object mask.

In summary, the amplitude complements the shortcomings of the phase. The advantages of fusing the two signals from low-frequency to high-frequency are as follows: Since the low-frequency THz signal provides precise depth (the thickness of an object) and fine edge/contour information in the phase and amplitude, respectively, they together better delineate and restore the object. In contrast, the high-

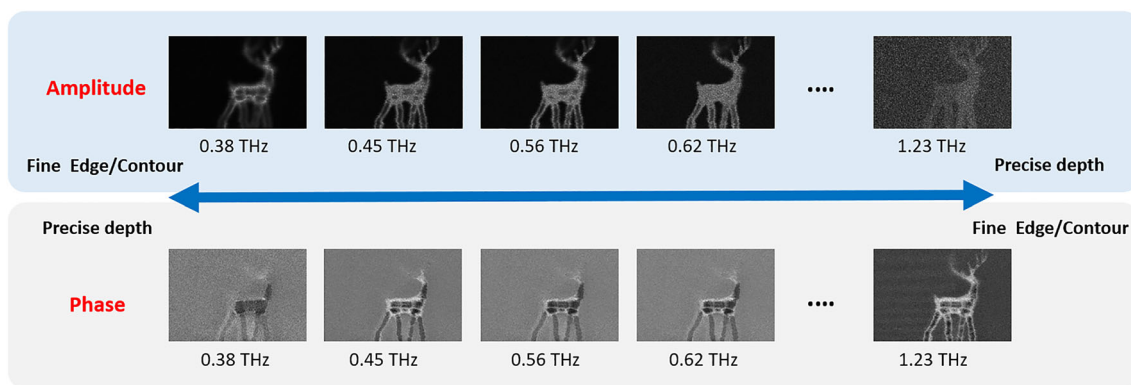


Fig. 3 Illustration of THz multi-spectral amplitude and phase images measured from **Deer**

frequency feature maps of amplitude and phase respectively provide better edges/contours and precise position information, thereby constituting a better object mask from the complementary features. With these multi-spectral properties of THz images, we can extract rich information from a wide spectral range in the frequency domain to simultaneously restore the 2D THz images without any additional computational cost or equipment, which is beneficial for the further development of THz imaging.

4 THz Tomographic Imaging

4.1 Overview

As different EM bands interact with objects differently, THz waves can partially penetrate through various optically opaque materials and carry hidden material tomographic information along the traveling path. This unique feature provides a new approach to visualizing the essence of 3D objects, which other imaging modalities cannot achieve. Although existing deep neural networks can learn spatio-spectral information from a considerable amount of spectral cube data, as mentioned above, directly learning from the **full spectral information** is not appropriate for learning THz image restoration models since the full spectral bands of THz signals involve diverse characteristics of materials, noises, and scattered signal, which causes difficulties in model training.

To address this problem, our work is based on extracting **complementary information** from both the amplitude and phase of a THz signal. In addition, for tomographic reconstruction, we capture multi-view images of an object with overlapping contents from different view angles. As a result, the correspondences existing between neighboring views offer useful aid in enhancing the restoration qualities of individual corrupted views. In summary, we devise a novel multi-view Subspace-Attention-guided Restoration Network ($SARNet_{MV}$), as shown in Fig. 4, to capture intra-

view complementary spectral characteristics of materials and inter-view correspondences from neighboring views to restore corrupted 2D THz images effectively. The key idea of $SARNet_{MV}$ is to fuse spatio-spectral features with different characteristics on a common ground via deriving a shared latent subspace and discovering the wide-range dependencies between the amplitude and phase images to guide the feature fusion. To this end, $SARNet_{MV}$ is a two-stage multi-view version based on single-view $SARNet$. In the first single-view stage of $SARNet_{MV}$, as shown in Fig. 4, all corrupted views are first restored by $SARNet$ individually. Then, in the second multi-view stage, we first concatenate and fuse the feature tensors of three restored neighboring views via a feature fusion module, and then feed the fused multi-view feature into the same $SARNet$ to obtain the final restored view. The design of $SARNet_{MV}$ is detailed in Sect. 4.2.

4.2 Network Architecture

On top of U-Net (Ronneberger et al., 2015), the architecture of $SARNet$ is depicted in Fig. 5. Specifically, $SARNet$ is composed of an encoder (spectral-fusion module) with 5 branches of different scales (from the finest to the coarsest) and a decoder (channel-fusion module) with 5 corresponding scale branches. Each scale branch of the encoder involves a Subspace-Attention-guided Fusion module (SAFM), a convolution block (Conv-block), and a down-sampler, except for the finest-scale branch that does not employ SAFM. To restore a specific view, the encoder of $SARNet$ takes the feature tensor of this view's Time-max image (the first stage) or a fused image of three restored neighboring views centered at the current view (the second stage) as the input of the finest-scale branch. To extract and fuse multi-spectral features of both amplitude and phase in a multi-scale manner, the encoder also receives to its second to fifth scale branches 24 images of additional predominant spectral frequencies extracted from the THz signal of the current view, where each branch takes 6 images of different spectral bands

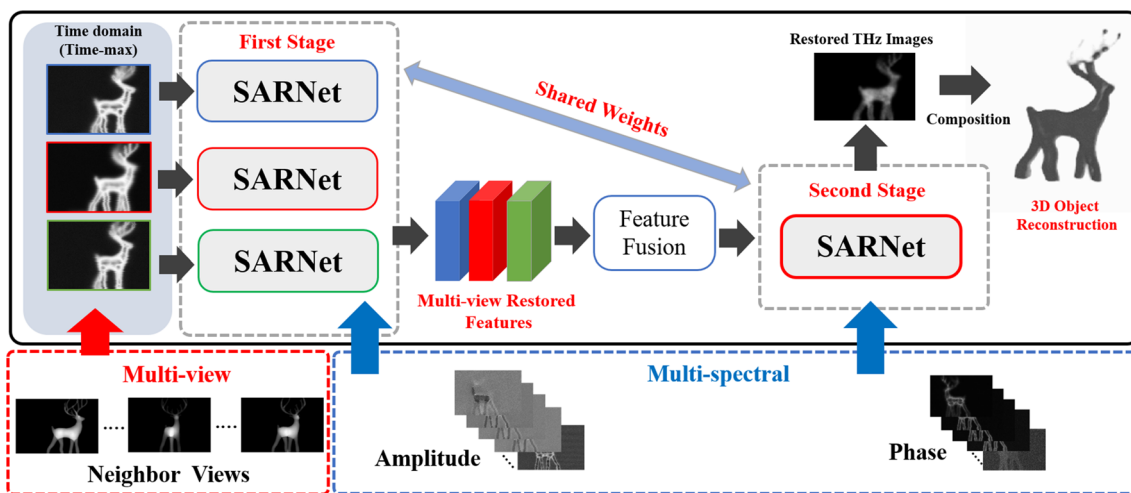


Fig. 4 Illustration of THz 3D tomographic imaging based on SARNet

(i.e., 3 amplitude bands and 3 corresponding phase bands) to extract learnable features from these spectral bands. To reduce the number of model parameters, these 24 amplitude and phase images (from low to high frequencies) are down-sampled to 4 different spatial scales and fed into the second to fifth scale branches in a fine-to-coarse manner as illustrated in Fig. 5. We then fuse the multi-spectral amplitude and phase feature maps in each scale via the proposed SAFM that learns a common latent subspace shared between the amplitude and phase features to facilitate associating the self-attention-guided wide-range amplitude-phase dependencies. Projected into the shared latent subspace, the spectral features of amplitude and phase components, along with the down-sampled features of the upper layer, can then be properly fused together on common ground in a fine-to-coarse fashion to derive the final latent code.

The Conv-block(L) contains two stacks of $L \times L$ convolution, batch normalization, and ReLU operations. Because the properties of the spectral bands of amplitude and phase can be significantly different, we partly use $L = 1$ to learn the best linear combination of multi-spectral features to avoid noise confusion and reduce the number of model parameters. The up-sampler and down-sampler perform $2 \times$ and $\frac{1}{2} \times$ scaling, respectively. The skip connections (SC) directly pass the feature maps of different spatial scales from individual encoder branches to the Channel Attention Modules (CAMs) of their corresponding branches of the decoder. The details of SAFM and CAM are elaborated on later.

In the decoder path, each scale branch for channel fusion involves an up-sampler, a CAM, and a Conv-block. The Conv-block has the same functional blocks as that in the encoder. Each decoding branch receives a “shallower-layer” feature map from the corresponding encoding branch via the skip-connection shortcut and concatenates the feature map with the upsampled version of the decoded “deeper-layer”

feature map from its coarser-scale branch. Besides, the concatenated feature map is then processed by CAM to capture the cross-channel interaction to complement the local region for restoration.

Note, a finer-scale branch of SARNet extracts shallower-layer features that tend to capture low-level features, such as colors and edges. To complement the Time-max image for restoration, we feed additional amplitude and phase images of low to high spectral bands into the fine- to coarse-scale branches of SARNet. Since the spectral bands of THz amplitude and phase offer complementary information, as mentioned in Sect. 4.1, besides the Time-max image SARNet also extracts multi-scale features from the amplitude and phase images of 12 selected THz spectral bands, which are then fused by the proposed SAFM.

4.3 Intra-view Feature Fusion of SARNet

4.3.1 Subspace-Attention-Guided Fusion Module

How to properly fuse the spectral features of THz amplitude and phase are, however, not trivial, as their characteristics can be significantly different. To address the problem, inspired by Cheng et al. (2021) and Zhang et al. (2019), we propose SAFM shown in Fig. 6 to fuse multi-spectral relevant features on common ground.

Let $\mathbf{X}_{in}^A, \mathbf{X}_{in}^P \in \mathbb{R}^{H \times W \times 3}$ denote the spectral bands of the THz amplitude and phase, respectively. The Conv-block $f_C(\cdot)$ extracts two intermediate feature maps $f_C(\mathbf{X}_{in}^A), f_C(\mathbf{X}_{in}^P) \in \mathbb{R}^{H \times W \times C_1}$ from \mathbf{X}_{in}^A and \mathbf{X}_{in}^P , respectively. As a result, we then derive the K shared basis vectors $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K]$ from $f_C(\mathbf{X}_{in}^A)$ and $f_C(\mathbf{X}_{in}^P)$, where $\mathbf{V} \in \mathbb{R}^{N \times K}$, $N = HW$ denotes the dimension of each basis vector, and K is the rank of the shared subspace. The basis set

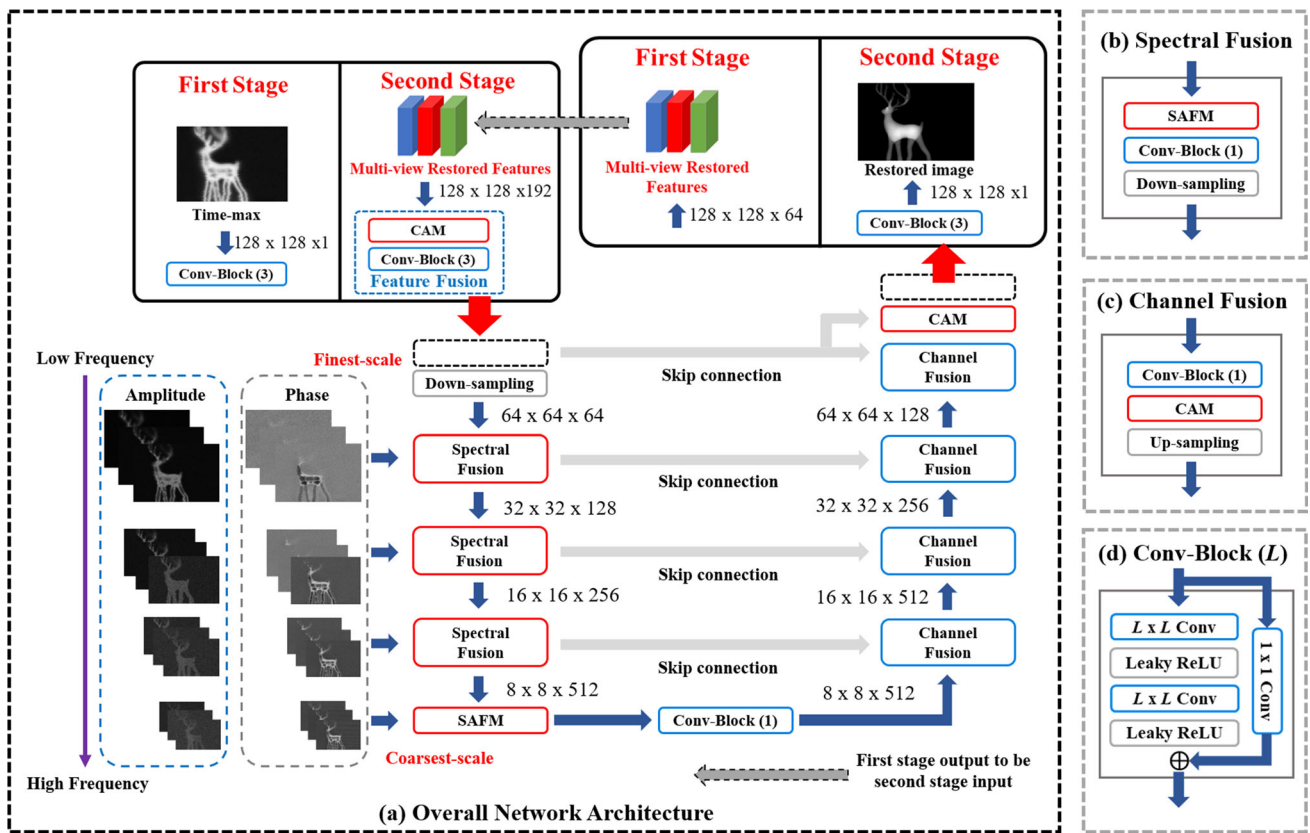
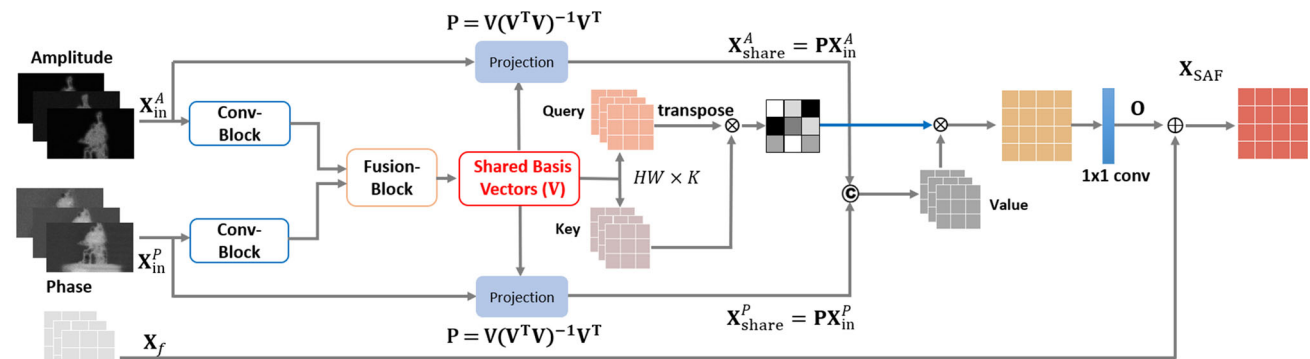


Fig. 5 a Overall network architecture of SARNet_{MV} consisting of five scale-branches, where the finest-scale scale takes the feature tensor of one view’s Time-max image (the first stage) or the fused image of three restored views (the second stage) as input. Additionally, each of the second to fifth takes 6 images of spectral frequencies (i.e., 3 ampli-

tude bands and 3 phase bands) as inputs. The three gray blocks show the detailed structures of **b** Spectral Fusion, **c** Channel Fusion, and **d** Conv-Block. The two black blocks indicate the input and output in the first and second stages, respectively



Down-sampled feature maps from the finer scale

Fig. 6 Block diagram of Subspace-Attention-guided Fusion Module (SAFM). SAFM first projects the different-band amplitude and phase features into a common latent subspace and then finds the wide-range

dependencies among the projected features via self-attention guidance. As a result, attention-aware features are fused on common ground

of the shared common subspace is expressed as

$$\mathbf{V} = f_F(f_C(\mathbf{X}_{in}^A), f_C(\mathbf{X}_{in}^P)), \tag{2}$$

where we first concatenate the two feature maps in the channel dimension and then feed the concatenated feature into the fusion block $f_F(\cdot)$. The structure of the fusion block is the same as that of the Conv-block with K output channels as indicated in the red block in Fig. 6. The weights of the fusion block are learned in the end-to-end training stage. The shared latent subspace learning mainly serves two purposes: (1) learning common latent representations between the THz amplitude and phase bands, and (2) learning the subspace projection matrix to project the amplitude and phase features into the shared subspace such that they can be analyzed on a common ground. These both help identify wide-range dependencies of amplitude and phase features for feature fusion.

To find wide-range dependencies between the amplitude and phase features on common ground, we utilize the orthogonal projection matrix \mathbf{V} in (2) to estimate the self-attentions in the shared feature subspace as

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^N \exp(s_{ij})}, \quad s_{ij} = \mathbf{v}_i^T \mathbf{v}_j \tag{3}$$

where $\beta_{j,i}$ represents the model attention in the i -th location of the j -th region.

The orthogonal projection matrix \mathbf{P} is derived from the subspace basis \mathbf{V} as follows (Meyer, 2000):

$$\mathbf{P} = \mathbf{V}(\mathbf{V}^T \mathbf{V})^{-1} \mathbf{V}^T \tag{4}$$

where $(\mathbf{V}^T \mathbf{V})^{-1}$ is the normalization term to make the basis vectors orthogonal to each other during the basis generation process.

As a result, the output of the self-attention mechanism becomes

$$\mathbf{o}_j = \left(\sum_{i=1}^N \beta_{j,i} \mathbf{s}_i \right), \quad \mathbf{s}_i = \text{Concate}(\mathbf{P}\mathbf{X}_{in}^A, \mathbf{P}\mathbf{X}_{in}^P) \tag{5}$$

where the key of $\mathbf{s}_i \in \mathbb{R}^{HW \times 6}$ is obtained by concatenating the two feature maps $\mathbf{P}\mathbf{X}_{in}^A$ and $\mathbf{P}\mathbf{X}_{in}^P$ projected by orthogonal projection matrix $\mathbf{P} \in \mathbb{R}^{HW \times HW}$, and \mathbf{X}_{in}^A and \mathbf{X}_{in}^P are reshaped to $HW \times 3$. Since the operations are purely linear with some proper reshaping, they are differentiable.

Finally, we further combine cross-scale features of the self-attention output by adding the down-sampled feature map \mathbf{X}_f from the finer scale as

$$\mathbf{Y}_{out}^{SAF} = f_s(\mathbf{o}) + \mathbf{X}_f \tag{6}$$

where f_s is a 1×1 convolution to keep the channel number consistent with \mathbf{X}_f .

4.3.2 Channel Attention Module

To fuse multi-scale features from different spectral bands in the channel dimension, we incorporate the efficient channel attention mechanism proposed in Qin et al. (2020) in the decoder path of SARNet as shown in Fig. 7. In each decoding branch, the original U-Net directly concatenates the up-sampled feature from the coarser scale with the feature from the corresponding encoding branch via the skip-connection shortcut, and then fuses the intermediate features from different layers by convolutions. This, however, leads to poor image restoration performances in local regions such as incorrect object thickness or details. To address this problem, we propose a channel attention module (CAM) that adopts full channel attention in the dimensionality reduction operation by concatenating two channel attention groups. CAM first performs global average pooling to extract the global spatial information in each channel:

$$G_t = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_t(i, j) \tag{7}$$

where $X_t(i, j)$ denotes the t -th channel of X_t at position (i, j) obtained by concatenating the up-sampled feature map \mathbf{X}_c of the coarser-scale and the skip-connection feature map \mathbf{X}_s . The shape of G is from $C \times H \times W$ to $C \times 1 \times 1$.

We directly feed the result through two stacks of 1×1 convolutions, sigmoid, and ReLU activation function as:

$$\mathbf{w} = \sigma(\text{Conv}_{1 \times 1}(\delta(\text{Conv}_{1 \times 1}(G)))) \tag{8}$$

where $\text{Conv}_{1 \times 1}(\cdot)$ denotes a 1×1 convolution, σ is the sigmoid function, and δ is the ReLU function. In order to better restore a local region, we divide the weights \mathbf{w} of different channels into two groups $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2]$ corresponding to two different sets of input feature maps, respectively. Finally, we element-wise multiply the input X_c and X_s of the weights \mathbf{w} and add these two group features.

4.4 Inter-view Feature Fusion of SARNet_{MV}

After restoring individual views of an object with SARNet, we then perform multi-view feature fusion between neighboring views to further boost restoration performance. As shown in Fig. 8, besides intra-view multi-spectral features, the inter-view redundancies between neighboring views can also provide informative clues for restoring corrupted views. To leverage the inter-view information, as shown in Fig. 5, for the t -th corrupted view, we fuse its post-restoration feature

Fig. 7 Block diagram of Channel Attention Module (CAM)

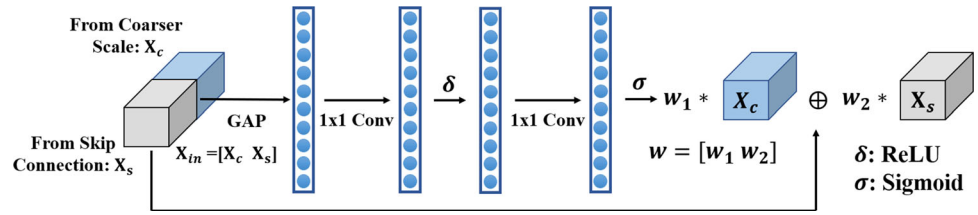
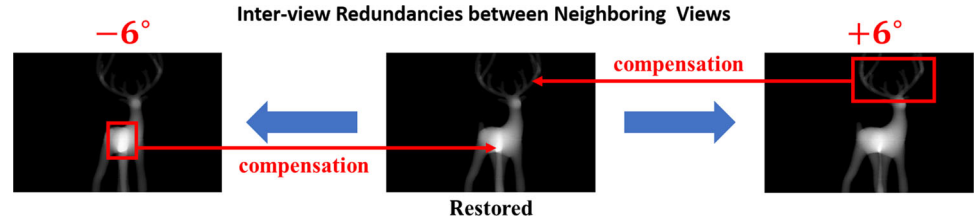


Fig. 8 Illustration of inter-view redundancies between neighboring views, where the redundancies existing in the two neighboring views offer useful clues for restoring the center view



tensor with those of its two closest views (i.e., the $(t - 1)$ -th and $(t + 1)$ -th views with a sampling step-size of 6°), all restored by the same SARNet model. To achieve inter-view feature fusion, we first concatenate the SARNet-restored feature tensors of three neighboring views centered at the t -th view, $\mathbf{X}_{in}^{(t-1)}$, $\mathbf{X}_{in}^{(t)}$, and $\mathbf{X}_{in}^{(t+1)}$, as follows:

$$\hat{\mathbf{X}}_{conc}^{(t)} = \text{Concat}(\mathbf{X}_{SAF}^{(t-1)}, \mathbf{X}_{SAF}^{(t)}, \mathbf{X}_{SAF}^{(t+1)}), \tag{9}$$

where $\text{Concat}(\cdot)$ denotes the concatenation operation and

$$\mathbf{X}_{SAF}^{(t)} = \text{SARNet} \left(\text{Conv}_{3 \times 3}(\mathbf{X}_{in}^{(t)}), \mathbb{W}(\mathbf{X}_{in}^{(t)}) \right), \tag{10}$$

where $\text{SARNet}(\cdot)$ denotes the restoration model, $\text{Conv}_{3 \times 3}(\cdot)$ denotes a 3×3 convolution, and $\mathbb{W}(\mathbf{X}_{in}^{(t)})$ denoted the set of 24 amplitude and phase spectral bands of $\mathbf{X}_{in}^{(t)}$ selected based on physics guidance.

The concatenated three-view feature tensor is then fused via the feature fusion block involving a CAM and a ConvBlock as follows:

$$\mathbf{X}_{MVF}^{(t)} = \text{Conv}_{3 \times 3} \left(\text{CAM} \left(\hat{\mathbf{X}}_{conc}^{(t)} \right) \right), \tag{11}$$

where $\mathbf{X}_{MVF}^{(t)}$ denotes the multi-view fused version of the t -th view, and $\text{CAM}(\cdot)$ is the channel attention module mentioned above.

Finally, as illustrated in Fig. we feed the fused three-view feature tensor \mathbf{X}_{MVF}^t into the finest-scale branch of SARNet along with the 24 amplitude and phase bands (i.e., the water absorption profile) associated with \mathbf{X}_{in}^t to obtain the final restoration result \mathbf{X}_{rec}^t as

$$\mathbf{X}_{rec}^{(t)} = \text{SARNet} \left(\mathbf{X}_{MVF}^{(t)}, \mathbb{W}(\mathbf{X}_{in}^{(t)}) \right). \tag{12}$$

4.5 Loss Function for THz Image Restoration

To effectively train SARNet, we employ the following mean squared error (MSE) loss function to measure the dissimilarity between the restored image \mathbf{X}_{rec} and its ground-truth \mathbf{X}_{GT} :

$$\mathcal{L}_{\text{MSE}}(\mathbf{X}_{GT}, \mathbf{X}_{rec}) = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W (\mathbf{X}_{GT}(i, j) - \mathbf{X}_{rec}(i, j))^2, \tag{13}$$

where H and W are the height and width of the image.

4.6 3D Tomographic Reconstruction

The 3D tomography of an object is reconstructed from the 60 2-D restored views of the object scanned from different angles. To reconstruct a 3-D tomography from the 60 2-D views, we directly utilize the inverse Radon transform to obtain the 3-D tomography, using methods like filtered back-projection (FPB) (Kak, 2001) or the simultaneous algebraic reconstruction technique (SART) (Recur et al., 2011).

5 Experimental Results

We conduct experiments to evaluate the effectiveness of SARNet against existing state-of-the-art restoration methods. We first present our THz-TDS system and measurement. Then, the details of the THz dataset and experiment settings. Finally, we evaluate the performances of SARNet and the competing methods on THz image restoration and tomographic reconstruction.

5.1 Proposed ASOPS THz-TDS System

Our in-house THz measurement system is an asynchronous optical sampling THz time-domain spectroscopy system

(ASOPS THz-TDS), which is composed of two asynchronous femtosecond lasers whose central wavelength are located at 1560 nm with tens of mW level, a pair of THz photoconductive antenna (THz PCA) source and detector, a linear and rotation motorized stage, four plane-convex THz lens with 50 mm focal length, a transimpedance amplifier (TIA), and a unit of data acquisition (DAQ) and processing (Janke et al., 2005). The repetition rates of the two asynchronous femtosecond lasers are 100 MHz and 100 MHz + 200 Hz, respectively. The sampling rate of DAQ is 20 MHz. With the configuration above, our ASOPS THz-TDS system delivers 0.1 ps temporal resolution and a THz frequency bandwidth of 5 THz. Additionally, our ASOPS THz-TDS system provides THz pulse signals with 41.7 dB dynamic range from 0.3 THz to 3 THz and 516 femtoseconds at full width at half maximum (FWHM). However, under the configuration above, the number of sampling points for a trace is approximately 100 K, consuming an extremely large transmission bandwidth. To address this limitation, only the 100-ps segment of the THz pulse signal is extracted. With the extracted segment of 100 ps, the frequency resolution is 10 GHz. Additionally, considering the minimum THz beam diameter of about 1.25 mm and the diffraction limitation, our THz system can provide spatial resolution in the scale of sub-millimeters.

5.2 Properties of THz Measurements

To retrieve the temporal-spatial-spectral information of each object voxel, our THz imaging experiment setup is based on a THz-TDS system as shown in Fig. 9. To demonstrate the THz penetrating capability, the measured object is first covered by a paper shield, which is highly transparent to THz but opaque in visible light. The covered object (e.g., a 3D printed deer covered by a paper shield) is placed on the rotation stage in the THz path between the THz source and detector of the THz-TDS system and is scanned by a raster scanning approach in 60 projection angles, as shown in Fig. 10.

Fig. 9 Illustration of THz data collection with our in-house THz-TDS tomographic imaging system

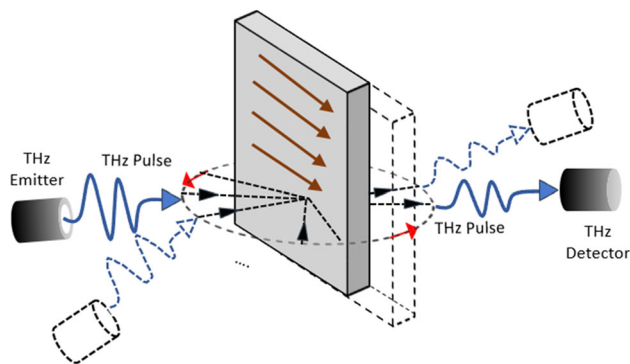
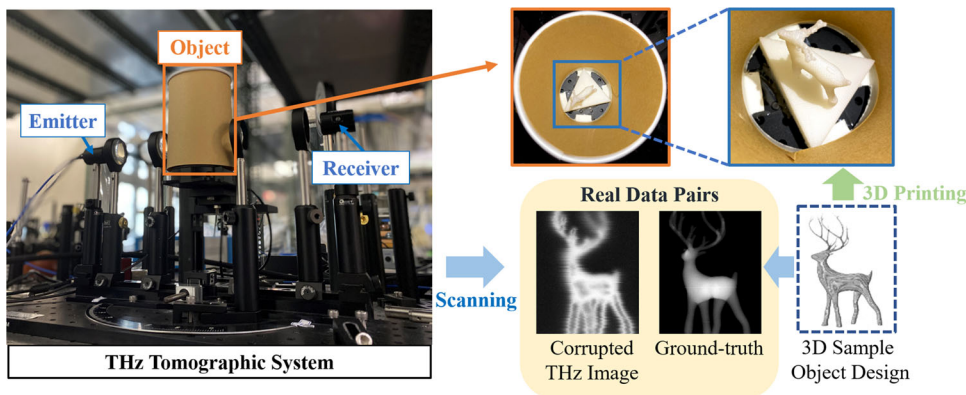


Fig. 10 The THz image formation system based on the raster scanning approach

During measuring, the THz-TDS system profiles each voxel’s THz temporal signal with 0.1 ps temporal resolution, whose amplitude corresponds to the strength of the THz electric field. With this scanning approach, a cube object of size 2 cm × 2 cm × 2 cm consumes about 1 minute for scanning a projected 2D image; thus, the cube will take about an hour for the 60 projection angles. Additionally, due to the limitation of the linear motorized stage, our measuring system can support an object size of about 6 cm at maximum. With our THz imaging experiment setup, the THz beam diameter varies with the THz propagation direction. As a consequence, the point-spread function of our system will vary with the geometry and location of the object. Therefore, the 2D projected images of the thickness-varying object could suffer from different levels of blurring effects in different pixels.

5.3 THz-TDS Image Dataset

As shown in Fig. 9, we prepare the sample objects by a Printech 3D printer, and use the material of high impact polystyrene (HIPS) for 3D-printing the objects.

The HIPS material is chosen since it can be used to quickly fabricate target objects by cost-efficient 3D printers, which can help evaluate a wide range of object geometries. Addi-

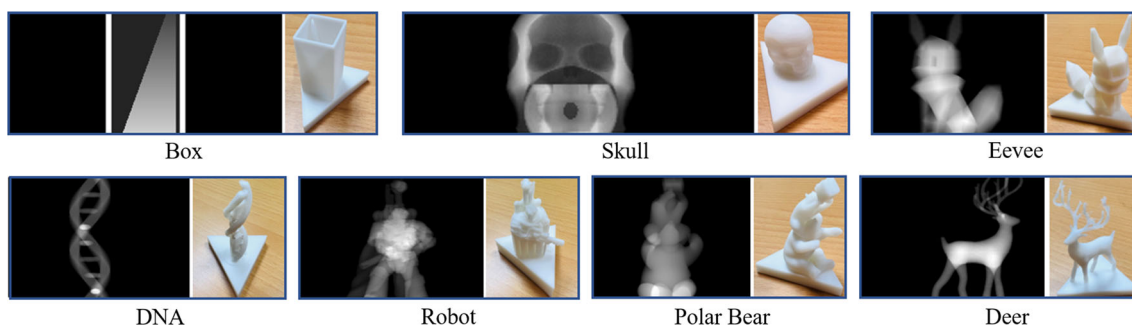


Fig. 11 Illustration of the ground-truths and photos of the seven 3D-printed HIPS objects used in our experiments. The left image of each object illustrates the ground-truth of one projection view and the right shows the photo of the HIPS object

tionally, the low absorption nature of HIPS in the THz range can prevent severe SNR degradation of detected THz signals while scanning objects. We then use our in-house ASOPS THz-TDS system (Janke et al., 2005) presented in Sect. 5.1 to measure the sample objects. Each sample object is placed on a motorized stage between the source and the receiver. With the help of the motorized stage, raster scans are performed on each object in multiple view angles. In the scanning phase, we scan the objects covering a rotational range of 180 degrees (step-size: 6 degrees between two neighboring views), a horizontal range of 72 mm (step-size: 0.25mm), and a variable vertical range corresponding to the object height (step-size: 0.25mm). In this way, we obtain 30 projections of each object, which are then augmented to 60 projections by horizontal flipping. The ground-truths of individual projections are obtained by taking the Radon transform of the 3D digital models defining the 3D object profiles for 3D printing in every view-angle. In addition to generating from digital models, the ground-truths can also be generated through precise 3D scanners.

We use markers to indicate the center of rotation so that we can align the ground-truths with the measured THz data. In this paper, totally seven sample objects are printed, measured, and aligned for evaluation.

5.4 Data Processing and Augmentation

In our experiments, we train the proposed multi-view $SARNet_{MV}$ model using the 2D THz images collected from our THz imaging system shown in Fig. 3. Figure 11 illustrates the photos of seven example objects along with their 2D ground-truths at certain projection angles. Each object consists of 60 projections and there are 420 2D THz images in total. In order to thoroughly evaluate the capacity of $SARNet_{MV}$, we adopt the leave-one-out strategy: using the data of 6 objects (i.e., 360 training images) as the training set, and that of the remaining object as the testing set. Due to the limited space, we only present part of the results in this section, and the complete results in the supplementary

material and our project site.¹ The THz-TDS image dataset can be found in the dataset site.² We will release our source code after the paper is accepted.

We also perform typical data augmentations to enrich the training set, including random rotating and flipping. Finally, the images are randomly cropped to 128×128 patches.

5.5 Experiment Settings

We initialize $SARNet_{MV}$ following the initialization method in He et al. (2015), and train it using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the initial learning rate to 10^{-4} and then decay the learning rate by 0.1 every 300 epochs. $SARNet$ converges after 1,000 epochs. For a fair comparison with the competing methods, we adopt their publicly released codes. All experiments were performed in a Python environment and Pytorch package running on a PC with Intel Core i7-10700 CPU 2.9 GHz and an Nvidia Titan 2080 Ti GPU.

5.6 Quantitative and Qualitative Evaluations

To the best of our knowledge, there is no method specially designed for restoring THz images besides `Time-max` (Hung & Yang, 2019b). Thus, we compare our $SARNet$ and $SARNet_{MV}$ models against several representative CNN-based image restoration models, including `DnCNN` (Zhang et al., 2017), `RED` (Mao et al., 2016), `NBNet` (Cheng et al., 2021), and `U-Net` (Ronneberger et al., 2015). We use the `time-max` images as the inputs and their corresponding round-truths as the target outputs to train these CNN-based image restoration models. Note that, these CNN-based restoration models do not utilize the prominent spectral information based on water absorption lines for restoring the THz `time-max` images. For quantitative quality assessment, we adopt two quality metrics for assessing the visual qualities of 2D view restora-

¹ Project site: https://github.com/wtnthu/THz_Tomography.

² Dataset site: https://github.com/wtnthu/THz_Data.

Table 2 Quantitative comparison of PSNR between restored 2D views and their ground-truth with different methods on **Deer**, **DNA**, **Box**, **Eevee**, **Polarbear**, **Robot**, and **Skull**. (\uparrow : higher is better)

Method	PSNR (dB) \uparrow						
	Deer	DNA	Box	Eevee	Polarbear	Robot	Skull
Time-max	12.42	12.07	11.97	11.20	11.21	11.37	10.69
DnCNN-S (Zhang et al., 2017)	19.94	23.95	19.13	19.69	19.44	19.72	17.33
RED (Mao et al., 2016)	19.30	24.17	20.18	19.97	19.17	19.76	16.28
NBNet (Cheng et al., 2021)	20.24	25.10	20.21	19.84	20.12	20.01	19.69
U-Net (Ronneberger et al., 2015)	19.84	24.15	19.77	19.95	19.09	18.80	17.49
SARNet (Ours)	22.98	26.05	22.67	20.87	21.42	22.66	22.48
SARNet _{MV} (Ours)	23.17	26.19	23.23	20.97	21.55	22.68	23.05

Table 3 Quantitative comparison of MSE between the cross-sections of a reconstructed 3D tomography and the corresponding ground-truths with different methods on **Deer**, **DNA**, **Box**, **Eevee**, **Polarbear**, **Robot**, and **Skull**. (\downarrow : lower is better)

Method	MSE \downarrow						
	Deer	DNA	Box	Eevee	Polarbear	Robot	Skull
Time-max	0.301	0.026	0.178	0.169	0.084	0.203	0.225
DnCNN-S (Zhang et al., 2017)	0.153	0.162	0.309	0.149	0.056	0.223	0.293
RED (Mao et al., 2016)	0.139	0.238	0.300	0.179	0.070	0.215	0.324
NBNet (Cheng et al., 2021)	0.240	0.184	0.305	0.134	0.088	0.128	0.138
U-Net (Ronneberger et al., 2015)	0.227	0.166	0.266	0.157	0.077	0.093	0.319
SARNet (Ours)	0.107	0.015	0.041	0.122	0.050	0.065	0.052
SARNet _{MV} (Ours)	0.091	0.013	0.030	0.105	0.038	0.059	0.049

tion and 3D tomographic reconstruction, respectively. and the reconstruction quality. The first metric is the Peak Signal-to-Noise Ratio (PSNR) for measuring the discrepancy between restored 2D views and their ground-truth as shown in Table 2. The second is the Mean-Square Error (MSE) between the cross-sections of a reconstructed 3D tomography and the corresponding ground-truths for assessing the 3D reconstruction accuracy as compared in Table 3. To further evaluate the 3D reconstruction accuracy of various models, as shown in Table 4, we also compare the average Intersection over Union (IoU), F-Score, and Chamfer distance performances by converting reconstructed 3D volumes into point-clouds (Xie et al., 2020).

Table 2 shows that our SARNet and SARNet_{MV} both significantly outperform the competing methods on all the seven sample objects in PSNR. Specifically, SARNet_{MV} outperforms Time-max (Hung & Yang, 2019b), U-Net (Ronneberger et al., 2015), and NBNet (Cheng et al., 2021) by 11.41 dB, 2.79 dB, and 2.23 dB, respectively, in average PSNR. In particular, even based on a simpler backbone U-Net, thanks to the good exploration of physics guidance, our proposed models significantly outperform the state-of-the-art restoration model NBNet especially on challenging objects like **Box** and **Skull**. With the aid of inter-view redundancies, multi-view SARNet_{MV} stably outperforms single-view SARNet and achieves notable 0.56 dB and 0.57 dB PSNR gains on **Box** and **Skull**. Similarly, in terms of 3D reconstruction accuracy, Table 3 demonstrates that our

models both stably achieve significantly lower average MSE of tomographic reconstruction than the competing methods on all seven objects. As for 3D shape reconstruction accuracy, Table 4 demonstrates that our models stably achieve significantly higher performances, in terms of average IoU, F-Score, and Chamfer distance of tomographic reconstruction, than the competing methods for all the seven objects.

For qualitative evaluation, Fig. 12 illustrates a few restored views for the seven sample objects, demonstrating that SARNet_{MV} can restore objects with much finer and smoother details (e.g., the antler and legs of **Deer**, the base pairs and shapes of **DNA** double-helix, the depth and shape of **Box**, the body and gun of **Robot**, and the correct depth and of **Skull**), the faithful thickness of material (e.g., the body and legs of **Deer**, and the correct edge thickness of **Box**), and fewer artifacts (e.g., holes and broken parts).

Our THz tomographic imaging system aims to reconstruct clear and faithful 3D object shapes. In our system, the tomography of an object is reconstructed from 60 views of 2D THz images of the object, each being restored by various image restoration models, via the inverse Radon transform. The paper shield region is cropped out to mitigate the evaluation bias caused by the simple geometry of the covered paper shield. Figure 13 illustrates the 3D reconstructions of the seven sample objects, showing that Time-max, U-Net tend to lose important object details such as holes in the deer's body with Time-max and the severely distorted antlers and legs with the three methods. In contrast, our method

Table 4 Quantitative comparison of average IoU, F-Score, and Chamfer distance between the 3D volume of a reconstructed 3D tomography and the corresponding ground-truths with different methods on **Deer**, **DNA**, **Box**, **Eevee**, **Polarbear**, **Robot**, and **Skull**. (\uparrow : higher is better and \downarrow : lower is better)

Method	IoU \uparrow						
	Deer	DNA	Box	Eevee	Polarbear	Robot	Skull
Time-max	0.247	0.427	0.106	0.323	0.482	0.041	0.385
DnCNN-S (Zhang et al., 2017)	0.179	0.136	0.096	0.509	0.353	0.260	0.158
RED (Mao et al., 2016)	0.386	0.323	0.257	0.359	0.433	0.142	0.175
NBNet (Cheng et al., 2021)	0.255	0.163	0.414	0.633	0.526	0.170	0.419
U-Net (Ronneberger et al., 2015)	0.400	0.427	0.117	0.423	0.539	0.290	0.286
SARNet (Ours)	0.502	0.515	0.418	0.702	0.550	0.434	0.407
SARNet _{MV} (Ours)	0.538	0.567	0.424	0.719	0.662	0.500	0.526
Method	F-Score \uparrow						
	Deer	DNA	Box	Eevee	Polarbear	Robot	Skull
Time-max	0.366	0.424	0.364	0.300	0.208	0.298	0.303
DnCNN-S (Zhang et al., 2017)	0.379	0.367	0.353	0.409	0.321	0.381	0.336
RED (Mao et al., 2016)	0.368	0.606	0.541	0.338	0.343	0.357	0.347
NBNet (Cheng et al., 2021)	0.476	0.278	0.506	0.346	0.268	0.314	0.381
U-Net (Ronneberger et al., 2015)	0.403	0.471	0.243	0.378	0.282	0.292	0.306
SARNet (Ours)	0.593	0.704	0.502	0.330	0.370	0.363	0.506
SARNet _{MV} (Ours)	0.605	0.715	0.574	0.410	0.391	0.400	0.613
Method	Chamfer distance \downarrow						
	Deer	DNA	Box	Eevee	Polarbear	Robot	Skull
Time-max	0.016	0.012	0.019	0.022	0.091	0.023	0.022
DnCNN-S (Zhang et al., 2017)	0.020	0.014	0.027	0.018	0.022	0.018	0.022
RED (Mao et al., 2016)	0.018	0.009	0.013	0.018	0.021	0.018	0.021
NBNet (Cheng et al., 2021)	0.016	0.020	0.013	0.020	0.025	0.021	0.021
U-Net (Ronneberger et al., 2015)	0.018	0.012	0.031	0.018	0.022	0.024	0.020
SARNet (Ours)	0.011	0.008	0.021	0.024	0.018	0.018	0.012
SARNet _{MV} (Ours)	0.011	0.008	0.011	0.016	0.016	0.018	0.010

reconstructs much clearer and more faithful 3D images with finer details, achieving by far the best 3D THz tomography reconstruction quality in the literature. Complete 3D reconstruction results are provided in the supplementary material.

Both the above quantitative and qualitative evaluations confirm a significant performance leap with SARNet_{MV} over the competing methods. Compared with our single-view restoration model (SARNet), the multi-view model SARNet_{MV} can restore finer local details such as the thickness of clear antlers, thinner edge of the box, and the gun in robot's hand. This also means that the inter-view redundancies between neighboring views are helpful in restoring local details, especially since our main task is to do 3D tomography. The correct thickness of the 2D image will directly affect the 3D tomography.

5.7 Ablation Studies

To verify the effectiveness of multi-spectral feature fusion, we evaluate the restoration performances with our SARNet_{MV}

under different settings in Table 5. The compared methods include (1) U-Net using a single channel of data (Time-max) without using features of multi-spectral bands; (2) U-Net+Amplitude w/o SAFM employing multi-band amplitude feature (without the SAFM mechanism) in each of the four spatial-scale branches, except for the finest scale (that accepts the Time-max image as the input), where 12 spectral bands of amplitude (3 bands/scale) are fed into the four spatial-scale branches with the assignment of the highest-frequency band to the coarsest scale, and vice versa; (3) U-Net+Phase w/o SAFM employing multi-spectral phase features with the same spectral arrangements as (2), and without the SAFM mechanism; (4) U-Net+Amplitude with SAFM utilizing subspace-attention-guided multi-spectral amplitude features with the same spectral arrangements as specified in (2); (5) U-Net+Phase with SAFM utilizing subspace-attention-guided multi-spectral phase features with the same spectral arrangements as in (2); (6) SARNet w/o SAFM concatenating multi-spectral amplitude and phase features (without SAFM) in each of the four spatial-scale

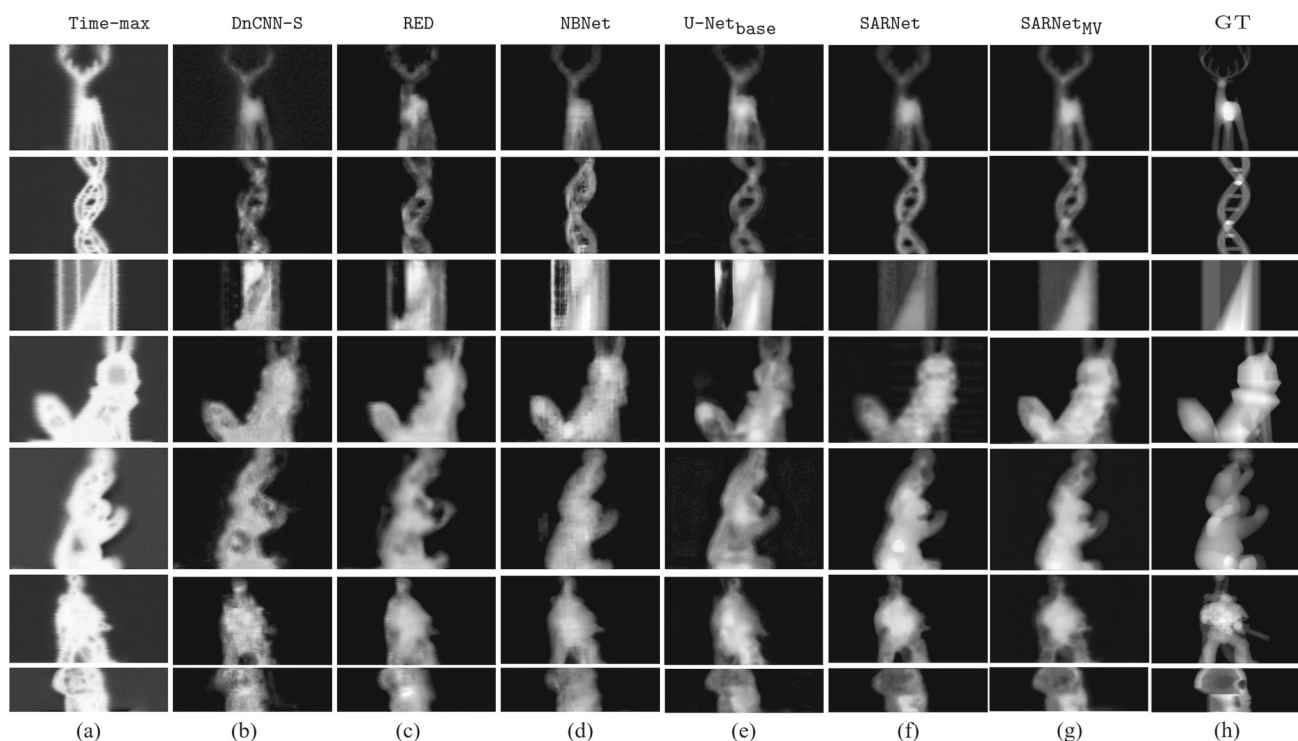


Fig. 12 Qualitative comparison of THz image restoration results for **Deer**, **DNA**, **Box**, **Eevee**, **Polarbear**, **Robot**, and **Skull** from left to right: **a** Time-max, **b** DnCNN-S (Zhang et al., 2017), **c** RED (Mao et

al., 2016), **d** NBNNet (Cheng et al., 2021), **e** U-Net_{base} (Ronneberger et al., 2015), **f** SARNet, **g** SARNet_{MV}, and **h** the ground-truth

branches, except for the finest scale (that accepts the Time-max image as the input), where totally 24 additional spectral bands of amplitude and phase (3 amplitude plus 3 phase bands for each scale) are fed into the four branches;(7) SARNet w/o Proj using SAFM to fuse intra-view multi-spectral amplitude and phase features without the aid of subspace projection; (8) SARNet_{MV} w/o SAFM employing multi-view fusion and multi-spectral amplitude and phase features with the same spectral arrangements as (6) but without subspace-attention guidance; (9) SARNet utilizing intra-view multi-spectral amplitude and phase features with subspace-attention guidance, but without utilizing the inter-view redundancies; (10) SARNet_{MV} utilizing full set of intra-view and inter-view features.

The results clearly demonstrate that the proposed SAFM can well fuse the spectral features of both amplitude and phase with different characteristics for THz image restoration. Specifically, employing additional multi-spectral features of either amplitude or phase as the input of the multi-scale branches in the network (i.e., U-Net+Amplitude w/o SAFM or U-Net+Phase w/o SAFM) can achieve performance improvement over U-Net. Combining both the amplitude and phase features without the proposed subspace-attention-guided fusion (i.e., SARNet w/o SAFM) does not outperform U-Net+Amplitude w/o SAFM and usually leads

to worse performances. The main reason is that the characteristics of the amplitude and phase features are too different to be fused to extract useful features with direct fusion methods. This motivates our subspace-attention-guided fusion scheme, which learns to effectively identify and fuse important and complementary features on common ground. The individual impacts of the subspace projection-guided fusion and the attention-guided fusion can be assessed by checking the performance differences among SARNet, SARNet w/o Proj, and SARNet w/o SAFM. Furthermore, the multi-view based SARNet_{MV} can further improve performance by utilizing additional inter-view redundancies, especially on objects with more details such as **Deer**, **Box**, and **Skull**. These results show that the proposed modules all stably achieve performance gains individually and collectively.

5.8 Model Complexity

Table 6 compares the model complexities of the six methods. When compared to the state-of-the-art method NBNNet (Cheng et al., 2021), and U-Net, our SARNet requires a much fewer number of parameters and GFLOPs. The run-time with SARNet is also less than NBNNet, but more than U-Net though. In contrast to SARNet, SARNet_{MV} achieves the best visual performance while introducing additional computa-

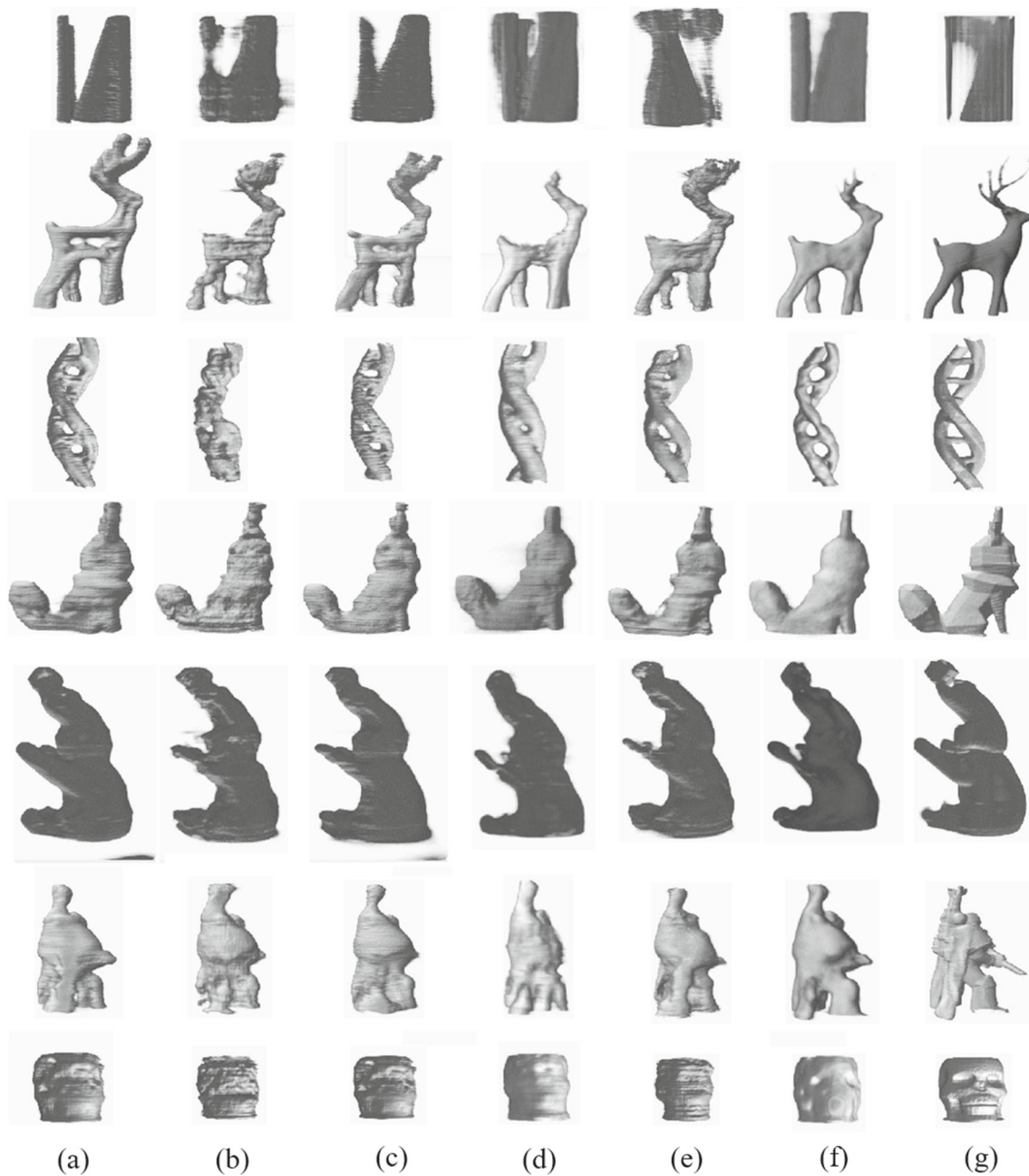


Fig. 13 Illustration of 3D tomographic reconstruction results on **Box**, **Deer**, **Dna**, **Eevee**, **Polarbear**, **Robot**, and **Skull** from left to right: **a** Time-max, **b** DnCNN-S (Zhang et al., 2017), **c** RED (Mao et al.,

2016), **d** NENet (Cheng et al., 2021), **e** U-Net_{base} (Ronneberger et al., 2015), **f** SARNet_{MV}, and **g** the ground-truth

tion and storage costs since it involves an additional stage of SARNet restoration, thereby doubling the computation. All the above comparisons demonstrate that both SARNet and SARNet_{MV} are promising solutions, considering their much better THz image restoration performances and reasonable computation and memory costs.

5.9 Limitations

SARNET uses multi-spectral amplitude/phase data to retrieve geometric information. Depending on the selected THz frequency bands and their SNR, the diffraction-limited system resolution can theoretically push down to 0.1mm. As water/metal are highly absorptive/reflective materials to the

Table 5 Ablation study in terms of PSNR of THz image restoration performances on **Deer, DNA, Box, Eevee, Polarbear, Robot, and Skull** with the different variants based on different settings. (\uparrow : higher is better)

Method	PSNR \uparrow						
	Deer	DNA	Box	Eevee	Polarbear	Robot	Skull
U-Net(Ronneberger et al., 2015)	19.84	24.15	19.77	19.95	19.09	18.80	17.49
U-Net+Amp w/oSAFM	22.05	25.84	20.32	20.21	20.48	20.63	20.70
U-Net+Phase w/oSAFM	21.14	24.98	20.42	20.26	20.15	20.58	21.36
U-Net+Amp w/ SAFM	20.97	26.00	21.83	20.22	20.30	21.11	20.18
U-Net+Phase w/ SAFM	22.66	25.52	21.65	20.63	20.18	21.50	21.42
SARNet w/o SAFM	21.44	25.78	20.00	20.32	20.44	21.12	21.18
SARNet w/o Proj	22.40	25.86	21.43	20.46	20.88	22.34	21.87
SARNet _{MV} w/o SAFM	22.49	25.78	22.10	19.91	20.96	21.75	22.47
SARNet (Ours)	22.98	26.05	22.67	20.87	21.42	22.66	22.48
SARNet _{MV} (Ours)	23.17	26.19	23.23	20.97	21.55	22.68	23.05

Table 6 Comparison of the model complexity (the numbers of Parameters, GFLOPs, and run-time) with different methods. Run-time are measured with the Nvidia Titan 2080 Ti

Method	Params (M)	GFLOPs	Run-time (ms)
DnCNN-S (Zhang et al., 2017)	0.55	4.55	6
RED (Mao et al., 2016)	0.66	1.36	4
NBNet(Cheng et al., 2021)	13.31	22.20	25
U-Net (Ronneberger et al., 2015)	9.5	3.88	11
SARNet (Ours)	3.0	1.91	19
SARNet _{MV} (Ours)	3.6	4.47	53

THz wave, our system is not applicable to the aqueous objects or objects hidden inside metallic packages.

Besides, limited by using a single THz source-detector pair, our THz-TDS system operates by a raster scanning approach. Although such a scanning approach makes it still far from real-time applications and is limited to static scenes, there are variants of THz-TDS systems that feature much shorter imaging time. For example, in Li and Jarrahi (2020), an N -pixel ($N = 63$) THz detector array is developed to offer N times faster image acquisition speed by spreading the THz light to the detector array.

6 Conclusions and Future Work

Aiming at making the invisible visible, we proposed a 3-D THz tomographic imaging system based on multi-view multi-scale spatio-spectral feature fusion. Considering the physical characteristics of THz waves passing through different materials, our THz imaging methods learn to extracting most predominant spectral features in different spatial scales for restoring corrupted THz images. The extracted multi-spectral features are then fused on a common ground by the proposed subspace-attention guided fusion and then used to restore THz images in a fine-to-coarse manner. As a result, the 3D tomography of an object can be reconstructed from the restored 2D THz images by inverse Radon transform for

object inspection and exploration. Besides intra-view fusion, we have also proposed an inter-view fusion approach to further improve the restoration/reconstruction performance. Our experimental results have confirmed a performance leap from the relevant state-of-the-art techniques in the area. We believe our findings in this work will shed on light on physics-guided THz computational imaging with advanced computer vision techniques.

As the THz computational imaging research in the computer vision community is still in its early stage, there are several possible directions worth further exploration. From the THz imaging quality point of view, an end-to-end learning framework for direct reconstruction of 3D geometry can avoid the artifacts caused by the typical tomographic reconstruction by the filtered backprojection of 2D projection views, thereby enhancing 3D reconstruction quality further. To this end, it would require to explore a newly designed learning framework involving network models, loss functions, and datasets. Moreover, incorporating the THz beam propagation 3D profile with the deconvolution techniques can further improve THz imaging quality. To extend the applications of THz imaging, by leveraging the prior knowledge of light-matter interaction in the THz range, the extension of THz computational imaging to functional imaging of multi-material objects can also be explored. Last but not least, integrating a massive THz detector array with the

THz-TDS system would pave the way to achieve real-time THz tomographic imaging.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11263-023-01812-y>.

Acknowledgements This work was financially supported in part by the National Science and Technology Council (NSTC), Taiwan, under Grants 111-2221-E-007-046-MY3, 111-2634-F-002-023, and 110-2636-E-007-017.

References

- Abbas, A., Abdelsamea, M., & Gaber, M. M. (2021). Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Applied Intelligence*, 51(2), 854–864.
- Abraham, E., Younus, A., Delagnes, T. C., & Mounaix, P. (2010). Non-invasive investigation of art paintings by terahertz imaging. *Applied Physics A*, 100(3), 585–590.
- Born, M., & Wolf, E. (2013). Principles of optics: Electromagnetic theory of propagation, interference and diffraction of light.
- Bowman, T., Chavez, T., Khan, K., Wu, J., Chakraborty, A., Rajaram, N., Bailey, K., & El-Shenawee, M. (2018). Pulsed terahertz imaging of breast cancer in freshly excised murine tumors. *Journal of Biomedical Optics*, 23(2), 026004.
- Calvin, Y., Shuting, F., Yiwen, S., & Emma, P.-M. (2012). The potential of terahertz imaging for cancer diagnosis: A review of investigations to date. *Quantitative Imaging in Medicine and Surgery*, 2(1), 33.
- Cao, J., Li, Y., Zhang, K., & Van Gool, L. (2021). Video super-resolution transformer. arXiv preprint [arXiv:2106.06847](https://arxiv.org/abs/2106.06847).
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Proceedings of European conference on computer vision* (pp. 213–229). Springer.
- Chapman, D., Homlinson, W., Johnston, R., Washburn, D., Pisano, E., Gmür, N., Zhong, Z., Menk, R., Arfelli, F., & Sayers, D. (1997). Diffraction enhanced X-ray imaging. *Journal Physics in Medicine & Biology*, 42(11), 2015.
- Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., & Gao, W. (2021). Pre-trained image processing transformer. In *Proceedings of IEEE/CVF conference on computer vision and pattern recognition* (pp. 12299–12310).
- Cheng, S., Wang, Y., Huang, H., Liu, D., Fan, H., & Liu, S. (2021). NBNNet: Noise basis learning for image denoising with subspace projection. In *Proceedings of IEEE/CVF international conference on computer vision and pattern recognition* (pp. 4896–4906).
- Clarke, L., Velthuisen, R., Camacho, M., Heine, J., Vaidyanathan, M., Hall, L., Thatcher, R., & Silbiger, M. (1995). MRI segmentation: Methods and applications. *Magnetic Resonance Imaging*, 13(3), 343–368.
- Cloetens, P., Barrett, R., Baruchel, J., Guigay, J.-P., & Schlenker, M. (1996). Phase objects in synchrotron radiation hard X-ray imaging. *Journal of Physics D: Applied Physics*, 29(1), 133.
- de Gonzalez, A. B., & Darby, S. (2004). Risk of cancer from diagnostic X-rays: Estimates for the UK and 14 other countries. *The Lancet*, 363(9406), 345–351.
- Dorney, T. D., Baraniuk, R. G., & Mittleman, D. M. (2001). Material parameter estimation with terahertz time-domain spectroscopy. *Journal of the Optical Society of America A*, 18(7), 1562–1571.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)
- Fitzgerald, R. (2000). Phase-sensitive X-ray imaging. *Physics Today*, 53(7), 23–26.
- Fukunaga, K. (2016). THz technology applied to cultural heritage in practice
- Geladi, P., Burger, J., & Lestander, T. (2004). Hyperspectral imaging: Calibration problems and solutions. *Chemometrics and Intelligent Laboratory Systems*, 72(2), 209–217.
- Hack, E., & Zolliker, P. (2014). Terahertz holography for imaging amplitude and phase objects. *Optics Express*, 22(13), 16079–16086.
- He, K., Zhang, X., Ren, S., Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of IEEE/CVF international conference on computer vision* (pp. 1026–1034).
- Hung, Y.-C., & Yang, S.-H. (2019a). Kernel size characterization for deep learning terahertz tomography (pp. 1–2).
- Hung, Y.-C., & Yang, S.-H. (2019b). Terahertz deep learning computed tomography. In *Proceedings of international infrared, millimeter, and terahertz waves* (pp. 1–2). IEEE.
- Janke, C., Först, M., Nagel, M., Kurz, H., & Bartels, A. (2005). Asynchronous optical sampling for high-speed characterization of integrated resonant terahertz sensors. *Optics Letters*, 30(11), 1405–1407.
- Jansen, C., Wietzke, S., Peters, O., Scheller, M., Vieweg, N., Salhi, M., Krumbholz, N., Jördens, C., Hochrein, T., & Koch, M. (2010). Terahertz imaging: Applications and perspectives. *Applied Optics*, 49(19), 48–57.
- Jin, K. H., McCann, M. T., Froustey, E., & Unser, M. (2017). Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9), 4509–4522.
- Kak, A. C. (2001). Algorithms for reconstruction with nondiffracting sources. *Principles of Computerized Tomographic Imaging*, 49–112.
- Kamruzzaman, M., ElMasry, G., Sun, D.-W., & Allen, P. (2011). Application of NIR hyperspectral imaging for discrimination of lamb muscles. *Journal of Food Engineering*, 104(3), 332–340.
- Kang, E., Min, J., & Ye, J. C. (2017). A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. *Journal of Medical Physics*, 44(10), 360–375.
- Kawase, K., Ogawa, Y., Watanabe, Y., & Inoue, H. (2003). Non-destructive terahertz imaging of illicit drugs using spectral fingerprints. *Optics Express*, 11(20), 2549–2554.
- Kim, J., Lim, H., Ahn, S.C., Lee, S. (2018). RGBD camera based material recognition via surface roughness estimation. In: *Proceedings of IEEE Winter Conference Applied Computer Vision* (pp. 1963–1971).
- Li, X., & Jarrahi, M. (2020). A 63-pixel plasmonic photoconductive terahertz focal-plane array. In *Proceedings of IEEE/MTT-S international microwave symposium (IMS)* (pp. 91–94).
- Liu, F., Jang, H., Kijowski, R., Bradshaw, T., & McMillan, A. B. (2018). Deep learning MR imaging-based attenuation correction for PET/MR imaging. *Radiology*, 286(2), 676–684.
- Ljubenic, M., Bazrafkan, S., Beenhouwer, J. D., & Sijbers, J. (2020). CNN-based deblurring of terahertz images (pp. 323–330).
- Mao, X., Shen, C., & Yang, Y.-B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Proceedings of the advances in neural information processing systems* (pp. 2802–2810).
- Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*. SIAM.
- Mittleman, D. M. (2018). Twenty years of terahertz imaging. *Optics Express*, 26(8), 9417–9431.
- Mittleman, D., Gupta, M., Neelamani, R., Baraniuk, R., Rudd, J., & Koch, M. (1999). Recent advances in terahertz imaging. *Applied Physics B*, 68(6), 1085–1094.

- Nunes-Pereira, E., Peixoto, H., Teixeira, J., & Santos, J. (2020). Polarization-coded material classification in automotive LIDAR aiming at safer autonomous driving implementations. *Applied Optics*, 59(8), 2530–2540.
- Ozdemir, A., & Polat, K. (2020). Deep learning applications for hyperspectral imaging: A systematic review. *Journal of the Institute of Electronics and Computer*, 2(1), 39–56.
- Peterson, J., Paerels, F., Kaastra, J., Arnaud, M., Reiprich, T., Fabian, A., Mushotzky, R., Jernigan, J., & Sakelliou, I. (2001). X-ray imaging-spectroscopy of Abell 1835. *Journal of Astronomy & Astrophysics*, 365(1), 104–109.
- Popescu, D. C., & Ellicar, A. D. (2010). Point spread function estimation for a terahertz imaging system. *EURASIP Journal on Advances in Signal Processing*, 2010(1), 575817.
- Popescu, D. C., Hellicar, A., & Li, Y. (2009). Phantom-based point spread function estimation for terahertz imaging system (pp. 629–639).
- Qin, X., Wang, X., Bai, Y., Xie, X., & Jia, H. (2020). FFA-Net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 11908–11915).
- Recur, B., Guillet, J.-P., Manek-Hönninger, I., Delagnes, J.-C., Benharbone, W., Desbarats, P., Domenger, J.-P., Canioni, L., & Mounaix, P. (2012). Propagation beam consideration for 3D THz computed tomography. *Optics Express*, 20(6), 5817–5829.
- Recur, B., Younus, A., Salort, S., Mounaix, P., Chassagne, B., Desbarats, P., Caumes, J., & Abraham, E. (2011). Investigation on reconstruction methods applied to 3D terahertz computed tomography. *Optics Express*, 19(6), 5105–5117.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of international conference on medical image computing and computer-assisted intervention* (pp. 234–241).
- Rotermund, H. H., Engel, W., Jakubith, S., Von Oertzen, A., & Ertl, G. (1991). Methods and application of UV photoelectron microscopy in heterogenous catalysis. *Ultramicroscopy*, 36(1–3), 164–172.
- Round, A. R., Wilkinson, S. J., Hall, C. J., Rogers, K. D., Glatter, O., Wess, T., & Ellis, I. O. (2005). A preliminary study of breast cancer diagnosis using laboratory based small angle x-ray scattering. *Physics in Medicine & Biology*, 50(17), 4159.
- Saeedkia, D. (2013). *Handbook of terahertz technology for imaging, sensing and communications* (pp. 542–578). Cambridge: Woodhead Publishing.
- Sakdinawat, A., & Attwood, D. (2010). Nanoscale X-ray imaging. *Nature Photonics*, 4(12), 840.
- Schultz, R., Nielsen, T., Zavaleta, R. J., Wyatt, R., & Garner, H. (2001). Hyperspectral imaging: A novel approach for microscopic analysis. *Cytometry*, 43(4), 239–247.
- Su, W.-T., Hung, Y.-C., Yu, P.-J., Lin, C.-W., & Yang, S.-H. (2023). Physics-guided terahertz computational imaging: A tutorial on state-of-the-art techniques. *IEEE Signal Processing Magazine*, 40(2), 32–45.
- Su, W.-T., Hung, Y.-C., Yu, P.-J., Yang, S.-H., & Lin, C.-W. (2022). Seeing through a black box: Toward high-quality terahertz tomographic imaging via multi-scale spatio-spectral image fusion. In *Proceedings of the European conference on computer vision*.
- Tuan, T. M., Fujita, H., Dey, N., Ashour, A. S., Ngoc, T. N., & Chu, D.-T. (2018). Dental diagnosis from X-ray images: An expert system based on fuzzy computing. *Biomedical Signal Processing and Control*, 39, 64–73.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of advances in neural information processing systems* (vol. 30).
- Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., & Li, H. (2022). Uformer: A general U-shaped transformer for image restoration. In *Proceedings of IEEE/CVF conference on computer vision and pattern recognition* (pp. 17683–17693).
- Wong, T. M., Kahl, M., & Bolívar, P. H., Kolb, A. (2019). Computational image enhancement for frequency modulated continuous wave (FMCW) THz image. *Journal of Infrared, Millimeter, and Terahertz Waves*, 40(7), 775–800.
- Wong, T. M., Kahl, M., Haring-Bolívar, P., Kolb, A., & Möller, M. (2019). Training auto-encoder-based optimizers for terahertz image reconstruction (pp. 93–106).
- Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint [arXiv:2006.03677](https://arxiv.org/abs/2006.03677).
- Xie, H., Yao, H., Zhang, S. P., Zhou, S. C., & Sun, W. X. (2020). Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12), 2919–2935.
- Xie, X. (2008). A review of recent advances in surface defect detection using texture analysis techniques. *ELCVIA: Electronic Letters on Computer Vision and Image Analysis*, 1–22.
- Yujiri, L., Shoucri, M., & Moffa, P. (2003). Passive millimeter wave imaging. *IEEE Microwave Magazine*, 4(3), 39–50.
- Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In *Proceedings of international conference on machine learning* (pp. 7354–7363).
- Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7), 3142–3155.
- Zhang, K., Zuo, W. M., & Zhang, L. (2018). FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9), 4608–4622.
- Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2020). Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., & Ren, J. (2019). Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2482–2491).
- Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, R. B., & Rosen, M. S. (2018). Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697), 487–492.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.