Seeing through a Black Box: Toward High-Quality Terahertz Imaging via Subspace-and-Attention Guided Restoration

Wen-Tai Su[1], Yi-Chun Hung[2], Po-Jen Yu[1], Shang-Hua Yang[1][0000−0002−5528−9281], and Chia-Wen Lin[1][0000−0002−9097−2318]

[1] Dept. EE, National Tsing Hua University, Hsinchu 300044, Taiwan
[2] Dept. ECE, University of California, Los Angeles, USA

Abstract. Terahertz (THz) 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 performances of existing restoration methods are highly constrained by the diffraction-limited THz signals. To address the problem, we propose a novel **S**ubspace-**A**ttention-guided **R**estoration **Net**work (`SARNet`) that fuses multi-spectral features of a THz image for effective restoration. To this end, `SARNet` uses multi-scale branches to extract spatio-spectral features of amplitude and phase which are then fused via shared subspace projection and attention guidance. Here, we experimentally construct a THz time-domain spectroscopy system covering a broad frequency range from 0.1 THz to 4 THz for building up temporal/spectral/spatial/phase/material THz database of hidden 3D objects. Complementary to a quantitative evaluation, we demonstrate the effectiveness of `SARNet` on 3D THz tomographic reconstruction applications.

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 [16,29,39]. X-ray imaging plays a crucial role in medical diagnosis, such as cancer, odontopathy, and COVID-19 symptom [1,30,35], 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 [38]. 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 [9]. On the other hand, although both non-ionizing and non-destructive, visible-light imaging cannot retrieve most optically opaque objects' interior information due to the highly absorptive and intense scattering behaviors between light and matter in the visible light band.

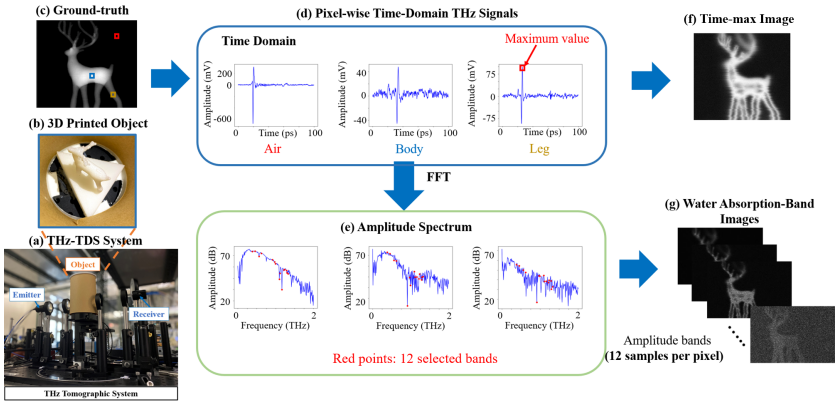

Fig. 1: THz data collection flow. (a) Our THz-TDS tomographic imaging system, (b) the 3D printed object, (c) the ground-truth of one projected view, (d) the time-domain THz signals of three different pixels (on the body and leg of the object and in the air), (e) the magnitude spectra of the three signals, (f) the time-max image (the maximums of each pixel's THz signal, (g) the images at the water-absorption frequencies.

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 [3,4].

THz radiation, in between microwave and infrared, has often been regarded as the last frontier of EM wave [31], 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 optically opaque materials, it carries hidden material tomographic information along the traveling path, making this approach a desired way to see through black boxes without damaging the exterior [21,13,22]. By utilizing light-matter interaction within the THz band, multifunctional tomographic information of 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 unique material interaction information in multi-dimensional domains — space, time, frequency, and phase, THz-TDS imaging has found applications in many emerging fields, including drug detection [18], industrial inspection, cultural heritage inspection [7], advanced material exploration [34], and cancer detection [2].

To retrieve temporal-spatio-spectral information of each object voxel, our THz imaging experiment setup is based on a THz-TDS system as shown in Fig. 1(a). Our measured object (a covered 3D printed deer, see Fig. 1(b)) is

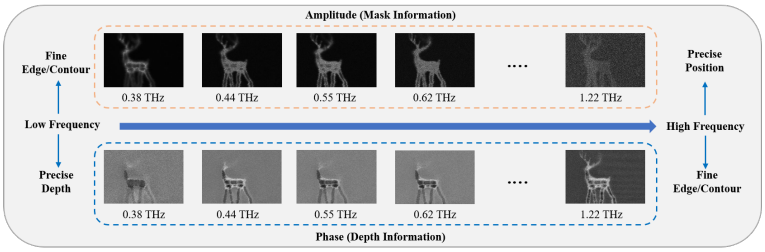

Fig. 2: THz multi-spectral amplitude and phase images measured from **Deer**.

placed on the rotation stage in the THz path between the THz source and detector of the THz-TDS system. During the scanning, the THz-TDS system profiles each voxel's THz temporal signal (Fig. 1(d)) with 0.1 ps temporal resolution, whose amplitude corresponds to the strength of THz electric field. 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`) is extracted as the feature for a voxel. The reconstructed image based on `Time-max` features can deliver great signal-to-noise ratio and a clear object contour. However, as shown in Fig. 1(f), 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 (Fig. 1(g)) of THz temporal signals to supplement the `Time-max` features since the voxel of the material behaviors are encoded in both the phase and amplitude of different frequency components, according to the Fresnel equation [6].

Due to the large number of spectral bands with measured THz image data, it is required to sample a subset of the spectral bands to reduce the training parameters. The THz beam is significantly attenuated at water absorption frequencies (*i.e.*, the valleys indicated in Fig. 1(e)). Thus, the reconstructed THz images based on water absorption lines offer worse details. Besides, our THz-TDS system offers more than 20 dB SNR in a frequency range of 0.3 THz–1.3 THz. Considering the water absorption in THz regime [36,33] and the superior SNR in the range of 0.3 THz–1.3 THz, we select 12 frequencies at 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. The spectral information including both amplitude and phase at the selected frequencies is extracted and used to restore clear 2D images. Fig. 2 depicts 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. Specifically, 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. 2). The phase also contains, however, a great variety of information of light-matter interaction that could

cause learning difficulty for the image restoration task. To address this issue, we utilize amplitude spectrum as complementary information. Although the attenuated amplitude spectrum cannot reflect comparable depth accuracy levels as phase spectrum, amplitude spectrum still present superior SNR and more faithful contours of a measured object. Besides, as complementary information to phase, the lower-frequency amplitude offers higher contrast, while the higher-frequency amplitude offers a better object mask.

In sum, the amplitude complements the shortcomings of phase. The advantages of fusing the two signals from low to high frequencies 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-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 restore 2D THz images without additional computational cost or equipment, which is beneficial for practical THz imaging systems.

We propose a **S**ubspace-**A**ttention-guided **R**estoration **Net** (`SARNet`) that fuses complementary THz amplitude and phase spectral features to supplement the `Time-max` image for restoring clear 2D images. To this end, `SARNet` learns common representations in a latent subspace shared between the amplitude and phase components, and then adopts a Self-Attention mechanism to learn the wide-range dependencies of the spectral features for guiding the restoration task. Finally, from clear 2D images restored from corrupted images of an object captured from different angles, we can reconstruct high-quality 3D tomography via inverse Radon transform. Our main contributions are summarized as follows:

- We merge the THz temporal-spatio-spectral data, physics-guided data-driven models, and material properties for high-precision THz tomographic imaging. The proposed `SARNet` has demonstrated the capability in extracting and fusing features from the light-matter interaction data in THz spectral regime, which inherently contains interior 3D object information and its material behaviors. Based on the designed architecture of `SARNet` on feature fusion, `SARNet` delivers state-of-the-art performance on THz image restoration.
- With our established THz tomography dataset, we provide comprehensive quantitative/qualitative analyses among `SARNet` and SOTAs. `SARNet` significantly outperforms `Time-max`, baseline U-Net, and multi-band U-Net by 11.17dB, 2.86dB, and 1.51dB in average PSNR.
- This proof-of-concept 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 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 network (CNNs) have proven to achieve the state-of-the-art performances in fundamental image restoration problem [41,19,43,42,28]. Several network models for image restoration were proposed, such as U-Net [28], hierarchical residual network [19] and residual dense network [43]. Notably, DnCNN [41] 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 [42] employs noise level maps as inputs and utilizes a single model to develop variants for solving problems with multiple noise levels. In [19] a very deep residual encoding-decoding (RED) architecture was proposed to solve the image restoration problem using skip connections. [43] proposed a residual dense network (RDN), which maximizes the reusability of features by using residual learning and dense connections. NBNet [5] 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 [24,25,37], and THz tomographic imaging works include [11,10].

2.2 Tomographic Reconstruction

Computer 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 [17], that adopts CNN to refine tomographic images. In [14], U-Net was used to refine image restoration with significantly improved performances. [44] 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 [32,23,8] 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.

3 Physics-Guided Terahertz Image Restoration

3.1 Overview

As different EM bands interact with objects differently, THz waves can partially penetrate through various optically opaque materials and carry hidden material

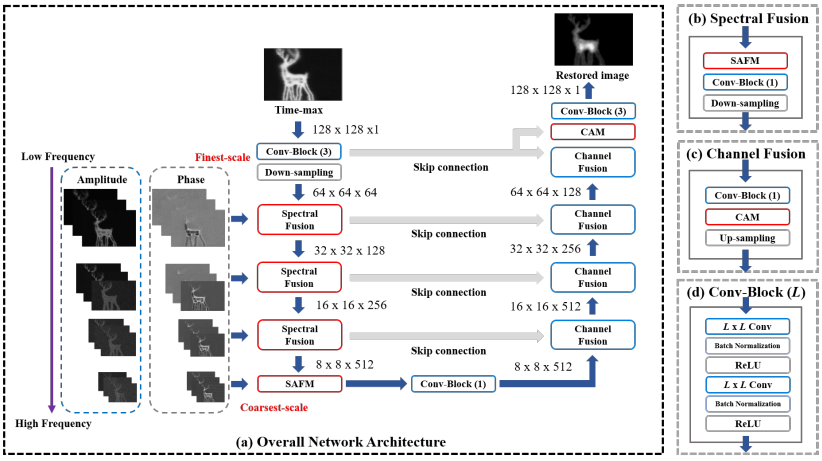

Fig. 3: (a) Overall network architecture of `SARNet` consisting of five scale-branches, where the finest-scale takes the feature tensor of one view's `Time-max` image as input. Additionally, each scale of the second to fifth takes 6 images of spectral frequencies (3 amplitude bands and 3 phase bands) as inputs. The three gray blocks show the detailed structures of (b) Spectral Fusion, (c) Channel Fusion, and (d) Conv-Block.

tomographic information along the traveling path. This unique feature provides a new approach to visualize 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, we found that directly learning from the **full spectral information** to restore THz images usually leads to an unsatisfactory performance. The main reason is that 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. That is, as illustrated in Fig. 2, in the low-frequency bands, the amplitude images delineate finer edges and object contours while the phase images offer relatively precise depths of object surfaces. In contrast, in the high-frequency bands, the amplitude images offer object mask information while the phase images delineate finer edges and object contours. Therefore, the amplitude and phase complement to each other in both the low and high frequency bands.

Motivated by the above findings, we devise a novel multi-scale `SARNet` to capture such complementary spectral characteristics of materials to restore damaged 2D THz images effectively. The key idea of `SARNet` is to fuse spatio-spectral features with different characteristics on a common ground via deriving the shared latent subspace and discovering the short/long-range dependencies between the amplitude and phase to guide the feature fusion. To this end, `SARNet` is based on U-Net[28] to perform feature extraction and fusion in a multi-scale manner.

3.2 Network Architecture

On top of U-Net [28], the architecture of `SARNet` is depicted in Fig. 3. 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 extract and fuse multi-spectral features of both amplitude and phase in a multi-scale manner, the encoder takes a THz 2D `Time-max` image as the input of the finest-scale branch as well as receives to its second to fifth scale branches 24 images of additional predominant spectral frequencies extracted from the THz signal, where each branch takes 6 images of different spectral bands (3 bands of amplitude and 3 bands of phase) 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 downsampled to 4 different resolutions and fed into the second to fifth scale branches in a fine-to-coarse manner as illustrated in Fig. 3. 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 short/long-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 a common ground in a fine-to-coarse fashion to obtain 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 will be elaborated later.

In the decoder path, each scale-branch for channel fusion involves a 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 which 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`.

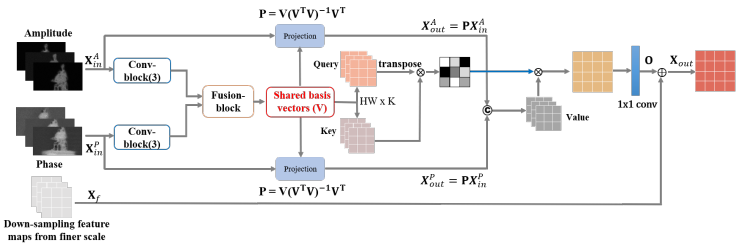

Fig. 4: Block diagram of Subspace-and-Attention guided Fusion Module (SAFM).

Since the spectral bands of THz amplitude and phase offer complementary information, as mentioned in Sec. 3.1, 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.

3.3 Subspace-Attention guided Fusion Module

How to properly fuse the spectral features of THz amplitude and phase are, however, not trivial, as their characteristics are very different. To address the problem, inspired by [5] and [40], we propose the SAFM as shown in Fig. 4.

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

$$\mathbf{V} = f_F(f_C(\mathbf{X}_{\text{in}}^{A}), f_C(\mathbf{X}_{\text{in}}^{P})), \tag{1}$$

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. 4. 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 the 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 a 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 a common ground, we utilize the orthogonal projection matrix $\mathbf{V}$ in (1) to estimate the self-attentions in the shared feature subspace as

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})}\ ,\ s_{ij} = \mathbf{v}_i^T\mathbf{v}_j \tag{2}$$

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

The projection matrix $\mathbf{P}$ is derived from the subspace basis $\mathbf{V}$ as follows [20]:

$$\mathbf{P} = \mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T \tag{3}$$

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

$$\mathbf{o}_j = \left(\sum_{i=1}^{N} \beta_{j,i}\mathbf{s}_i\right), \quad \mathbf{s}_i = \text{Concate}(\mathbf{P}\mathbf{X}_{\text{in}}^A, \mathbf{P}\mathbf{X}_{\text{in}}^P) \tag{4}$$

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

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

$$\mathbf{X}_{\mathbf{out}} = f_s(\mathbf{o}) + \mathbf{X}_f \tag{5}$$

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

3.4 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 [26] in the decoder path of `SARNet`. 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:

$$G_t = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_t(i,j) \tag{6}$$

where $X_t(i,j)$ is 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 1×1 convolution, sigmoid, and ReLU activation function as:

$$\mathbf{w} = \sigma\left(\mathrm{Conv}_{1\times1}\left(\delta\left(\mathrm{Conv}_{1\times1}(G)\right)\right)\right), \tag{7}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ 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.

3.5 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 $\mathbf{X}_{\mathrm{rec}}$ and its ground-truth $\mathbf{X}_{\mathrm{GT}}$:

$$\mathcal{L}_{\mathrm{MSE}}(\mathbf{X}_{\mathrm{GT}}, \mathbf{X}_{\mathrm{rec}}) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathbf{X}_{\mathrm{GT}}(i,j) - \mathbf{X}_{\mathrm{rec}}(i,j))^2, \tag{8}$$

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

3.6 3D Tomography Reconstruction

The 3D tomography of an object can then be reconstructed from the restored THz 2D images of the object scanned in different angles. To this end, we can directly apply the inverse Radon transform to obtain the 3D tomography, using methods like filtered back-projection [15] or the simultaneous algebraic reconstruction technique [27].

4 Experiments

We conduct experiments to evaluate the effectiveness of `SARNet` against existing state-of-the-art restoration methods. We first present our experiment settings and then evaluate the performances of `SARNet` and the competing methods on THz image restoration.

4.1 THz-TDS Image Dataset

As shown in Fig. 1, we prepare the sample objects by a Printech 3D printer and use the material of high impact polystyrene (HIPS) for 3D-printing the objects due to its high penetration of THz waves. We then use our in-house Asynchronous Optical Sampling (ASOPS) THz-TDS system [12] to measure the sample objects. Although the speed of our mechanical scanning stage limits the number and the size of the objects in the dataset, we carefully designed 7

Table 1: Quantitative comparison (PSNR and SSIM) of THz image restoration performances with different methods on seven test objects. (↑: higher is better)

Method	PSNR↑							SSIM↑						
	Deer	DNA	Box	Eevee	Bear	Robot	Skull	Deer	DNA	Box	Eevee	Bear	Robot	Skull
`Time-max`	12.42	12.07	11.97	11.20	11.21	11.37	10.69	0.05	0.05	0.14	0.14	0.12	0.08	0.09
`DnCNN-S` [41]	19.94	23.95	19.13	19.69	19.44	19.72	17.33	0.73	0.77	0.73	0.72	0.63	0.77	0.36
`RED` [19]	19.30	24.17	20.18	19.97	19.17	19.76	16.28	0.81	0.83	0.74	0.77	0.75	0.80	0.74
`NBNet` [5]	20.24	25.10	20.21	19.84	20.12	20.01	19.69	0.81	0.85	0.75	0.77	0.80	0.80	0.78
$\texttt{U-Net}_{\texttt{base}}$ [28]	19.84	24.15	19.77	19.95	19.09	18.80	17.49	0.55	0.78	0.77	0.76	0.56	0.76	0.51
$\texttt{U-Net}_{\texttt{MB}}$	22.46	25.05	20.81	20.34	19.86	20.64	19.43	0.76	0.73	0.78	0.76	0.78	0.79	0.78
`SARNet` (Ours)	**22.98**	**26.05**	**22.67**	**20.87**	**21.42**	**22.66**	**22.48**	**0.84**	**0.90**	**0.83**	**0.82**	**0.82**	**0.83**	**0.84**

objects to increase the dataset variety for the generalization to unseen objects. For example, the antler of Deer and the tilted cone of Box are designed for high spatial frequency and varying object thickness. 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), a horizontal range of 72mm (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 converting the original 3D printing files into image projections in every view-angle. We use markers to indicate the center of rotation to align the ground truths with the measured THz data. In this paper, a total of 7 objects are printed, measured, and aligned for evaluation.

4.2 Data Processing and Augmentation

In our experiments, we train `SARNet` using the 2D THz images collected from our THz imaging system shown in Fig. 2. The seven sample objects are consisting of 60 projections per object and 420 2D THz images in total. To evaluate the effectiveness of `SARNet`, we adopt the leave-one-out strategy: using the data of 6 objects 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. We will release our code (Link) and the THz image dataset (Link) after the work is accepted.

4.3 Quantitative Evaluations

To the best of our knowledge, there is no method specially designed for restoring THz images besides `Time-max`. Thus, we compare our method against several representative CNN-based image restoration models, including DnCNN [41], RED [19], and NBNet [5]. Moreover, we also compare two variants of U-Net [28]: baseline U-Net ($\texttt{U-Net}_{\texttt{base}}$) and multi-band U-Net ($\texttt{U-Net}_{\texttt{MB}}$). $\texttt{U-Net}_{\texttt{base}}$ extracts

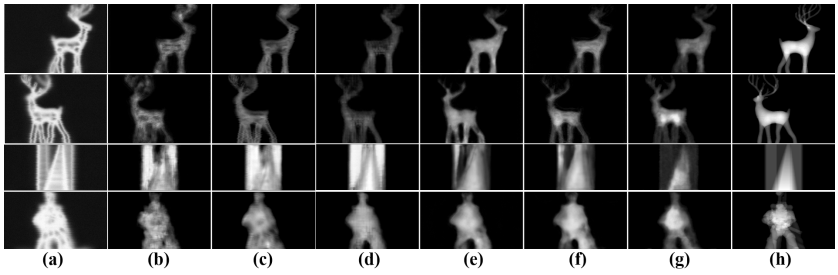

Fig. 5: Qualitative comparison of THz image restoration results for **Deer**, **Box**, and **Robot** from left to right: (a) `Time-max`, (b) `DnCNN-S` [41], (c) `RED` [19], (d) `NBNet` [5], (e) $\texttt{U-Net}_{\texttt{base}}$ [28], (f) $\texttt{U-Net}_{\texttt{MS}}$, (g) `SARNet`, and (h) the ground-truth.

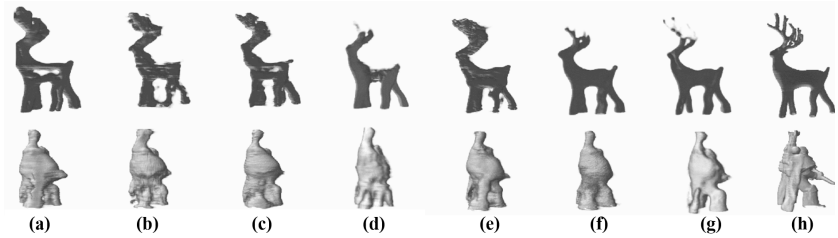

Fig. 6: Illustration of 3D tomographic reconstruction results on **Deer** and **Robot** from left to right: (a) `Time-max`, (b) `DnCNN-S` [41], (c) `RED` [19], (d) `NBNet` [5], (e) $\texttt{U-Net}_{\texttt{base}}$ [28], (f) $\texttt{U-Net}_{\texttt{MB}}$, (g) `SARNet`, and (h) the ground-truth.

image features in five different scales following the original setting in U-Net [28], whereas $\texttt{U-Net}_{\texttt{MB}}$ incorporates multi-spectral features by concatenating the features of `Time-max` image with additional 12 THz bands for amplitude as the input (i.e., $12 + 1$ channels) of the finest scale of U-Net. For objective quality assessment, we adopt two widely-used metrics including the Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) to respectively measure the pixel-level and structure-level similarities between a restored image and its ground-truth. To estiamte the 3D tomographic reconstruction, we adopt 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 2.

Table 1 shows that our `SARNet` significantly outperforms the competing methods on all sample objects in both metrics. Specifically, `SARNet` outperforms `Time-max`, baseline U-Net ($\texttt{U-Net}_{\texttt{base}}$), and the multi-band U-Net ($\texttt{U-Net}_{\texttt{MB}}$) by 11.17 dB, 2.86 dB, and 1.51 dB in average PSNR, and 0.744, 0.170, and 0.072 in average SSIM for 7 objects. Similarly, in terms of 3D reconstruction accuracy, Table 2 demonstrates that our models both stably achieve significantly lower average MSE of tomographic reconstruction than the competing methods on all the seven objects. For qualitative evaluation, Fig. 5 illustrates a few restored

Table 2: Quantitative comparison of MSE between the cross-sections of reconstructed 3D objects and the ground-truths with different methods on 7 objects. (↓: lower is better)

Method	MSE↓						
	Deer	DNA	Box	Eevee	Bear	Robot	Skull
`Time-max`	0.301	0.026	0.178	0.169	0.084	0.203	0.225
`DnCNN-S` [41]	0.153	0.162	0.309	0.149	0.056	0.223	0.293
`RED` [19]	0.139	0.238	0.300	0.179	0.070	0.215	0.324
`NBNet` [5]	0.240	0.184	0.305	0.134	0.088	0.128	0.138
`U-Net`$_{\texttt{base}}$ [28]	0.227	0.166	0.266	0.157	0.077	0.093	0.319
`U-Net`$_{\texttt{MB}}$	0.183	0.043	0.205	0.114	0.089	0.196	0.080
`SARNet` (Ours)	**0.107**	**0.015**	**0.041**	**0.105**	**0.050**	**0.065**	**0.052**

Table 3: Quantitative comparison (PSNR and SSIM) of THz image restoration performances on seven test objects with the different variants of `SARNet` based on different settings. (↑: higher is better)

Method	PSNR↑							SSIM↑						
	Deer	DNA	Box	Eevee	Bear	Robot	Skull	Deer	DNA	Box	Eevee	Bear	Robot	Skull
`U-Net`$_{\texttt{base}}$	19.84	25.63	19.77	19.95	19.09	18.80	10.69	0.55	0.78	0.77	0.76	0.56	0.76	0.51
Amp-Unet w/o SAFM	22.05	25.84	20.32	20.21	20.48	20.63	20.70	0.80	0.83	0.77	0.79	0.80	0.78	0.77
Phase-Unet w/o SAFM	21.14	24.98	20.42	20.26	20.15	20.58	21.36	0.82	0.72	0.78	0.78	0.81	0.74	0.75
Mix-Unet w/o SAFM	21.44	25.78	20.00	20.32	20.44	21.12	21.18	0.81	0.81	0.78	0.80	0.79	0.81	0.82
Amp-Unet w/ SAFM	20.97	26.00	21.83	20.22	20.30	21.11	20.18	0.84	0.90	0.78	0.80	0.79	0.83	0.79
Phase-Unet w/ SAFM	22.66	25.52	21.65	20.63	20.18	21.50	21.42	0.83	0.86	0.79	0.74	0.81	0.83	0.82
`SARNet` (Ours)	**22.98**	**26.05**	**22.67**	**20.87**	**21.42**	**22.66**	**22.48**	**0.84**	**0.90**	**0.83**	**0.82**	**0.82**	**0.83**	**0.84**

views for **Deer**, **Box**, and **Robot**, demonstrating that `SARNet` can restore objects with much finer and smoother details (e.g., the antler and legs of **Deer**, the depth and shape of **Box**, and the body of **Robot**), 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). Both the quantitative and qualitative evaluations confirm a significant performance leap with `SARNet` over the competing methods.

4.4 Ablation Studies

To verify the effectiveness of multi-spectral feature fusion, we evaluate the restoration performances with `SARNet` under different settings in Table 3. The compared methods include (1) `U-Net`$_{\texttt{base}}$ using a single channel of data (`Time-max`) without using features of multi-spectral bands; (2) **Amp-Unet 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) **Phase-Unet w/o SAFM** employing multi-spectral phase features with the same spectral arrangements as (2), and without the SAFM mechanism; (4) **Mix-Unet w/o SAFM** concatenating

multi-spectral amplitude and phase features (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 totally 24 additional spectral bands of amplitude and phase (3 amplitude plus 3 phase bands for each scale) are fed into the four branches; (5) **Amp-Unet with SAFM** utilizing attention-guided multi-spectral amplitude features with the same spectral arrangements as specified in (2); and (6) **Phase-Unet with SAFM** utilizing attention-guided multi-spectral phase features with the same spectral arrangements as in (2).

The results clearly demonstrate that the proposed SAFM can benefit fusing 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., Amp-Unet or Phase-Unet w/o SAFM) can achieve performance improvement over $\texttt{U-Net}_{\texttt{base}}$. Combining both the amplitude and phase features without the proposed subspace-and-attention guided fusion (i.e., **Mix-Unet w/o SAFM**) does not outperform **Amp-Unet 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-and-attention guided fusion scheme, that learns to effectively identify and fuse important and complementary features on a common ground.

4.5 3D Tomography Reconstruction

Our goal is to reconstruct clear and faithful 3D object shapes through our THz tomographic imaging system. In our system, the tomography of an object is reconstructed from 60 views of 2D THz images of the object, each being restored by `SARNet`, via the inverse Radon transform. Fig. 6 illustrates the 3D reconstructions of our reconstruct results much clearer and more faithful 3D images with finer details such as the thickness of body and clear antlers of of **Deer** and the gun in **Robot**'s hand, achieving by far the best 3D THz tomography reconstruction quality in the literature. Complete 3D reconstruction results are provided in the supplementary material.

5 Conclusion

We proposed a 3D THz imaging system that is the first to merge THz spatio-spectral data, data-driven models, and material properties to restore corrupted THz images. Based on the physical characteristics of THz waves passing through different materials, our `SARNet` efficiently fuses 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 to guide the feature fusion for boosting restoration performance. Our 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 stimulate further applicable research for THz imaging with advanced computer vision techniques.

References

1. Abbas, A., Abdelsamea, M., Gaber, M.: Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network. Appl. Intell. **51**(2), 854–864 (2021)
2. Bowman, T., Chavez, T., Khan, K., Wu, J., Chakraborty, A., Rajaram, N., Bailey, K., El-Shenawee, M.: Pulsed terahertz imaging of breast cancer in freshly excised murine tumors. J. Biomedical optics **23**(2), 026004 (2018)
3. braham, E., Younus, A., Delagnes, T.C., Mounaix, P.: Non-invasive investigation of art paintings by terahertz imaging. Applied Physics A **100**(3), 585–590 (2010)
4. Calvin, Y., Shuting, F., Yiwen, S., Emma, P.M.: The potential of terahertz imaging for cancer diagnosis: A review of investigations to date. Quantitative imaging in medicine and surgery **2**(1), 33 (2012)
5. Cheng, S., Wang, Y., Huang, H., Liu, D., Fan, H., Liu, S.: NBNet: Noise basis learning for image denoising with subspace projection. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. pp. 4896–4906 (2021)
6. Dorney, T.D., Baraniuk, R.G., Mittleman, D.M.: Material parameter estimation with terahertz time-domain spectroscopy. JOSA A **18**(7), 1562–1571 (2001)
7. Fukunaga, K.: Thz technology applied to cultural heritage in practice. Springer (2016)
8. Geladi, P., Burger, J., Lestander, T.: Hyperspectral imaging: calibration problems and solutions. Chemometrics and intelligent laboratory systems **72**(2), 209–217 (2004)
9. de Gonzalez, A.B., Darby, S.: Risk of cancer from diagnostic x-rays: estimates for the uk and 14 other countries. The Lancet **363**(9406), 345–351 (2004)
10. Hung, Y.C., Yang, S.H.: Kernel size characterization for deep learning terahertz tomography. In: Proc. IEEE Int. Conf. Infrared, Millimeter, and Terahertz Waves (IRMMW-THz). pp. 1–2 (2019)
11. Hung, Y.C., Yang, S.H.: Terahertz deep learning computed tomography. In: Proc. Int. Infrad. Milli. THz. Wav. pp. 1–2. IEEE (2019)
12. Janke, C., Först, M., Nagel, M., Kurz, H., Bartels, A.: Asynchronous optical sampling for high-speed characterization of integrated resonant terahertz sensors. Optics Lett. **30**(11), 1405–1407 (2005)
13. Jansen, C., Wietzke, S., Peters, O., Scheller, M., Vieweg, N., Salhi, M., Krumbholz, N., Jördens, C., Hochrein, T., Koch, M.: Terahertz imaging: applications and perspectives. Appl. Optics **49**(19), E48–E57 (2010)
14. Jin, K.H., McCann, M.T., Froustey, E., Unser, M.: Deep convolutional neural network for inverse problems in imaging. IEEE Trans. Image Process. **26**(9), 4509–4522 (2017)
15. Kak, A.C.: Algorithms for reconstruction with nondiffracting sources. Principles of computerized tomographic imaging pp. 49–112 (2001)
16. Kamruzzaman, M., ElMasry, G., Sun, D.W., Allen, P.: Application of nir hyperspectral imaging for discrimination of lamb muscles. J. Food Engineer. **104**(3), 332–340 (2011)
17. Kang, E., Min, J., Ye, J.C.: A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction. J. Medical physics **44**(10), e360–e375 (2017)
18. Kawase, K., Ogawa, Y., Watanabe, Y., Inoue, H.: Non-destructive terahertz imaging of illicit drugs using spectral fingerprints. Optics Express **11**(20), 2549–2554 (2003)

19. Mao, X., Shen, C., Yang, Y.B.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In: Proc. Adv. Neural Inf. Process. Syst. p. 2802–2810 (2016)
20. Meyer, C.D.: Matrix Analysis and Applied Linear Algebra. SIAM (2000)
21. Mittleman, D., Gupta, M., Neelamani, R., Baraniuk, R., Rudd, J., Koch, M.: Recent advances in terahertz imaging. Applied Physics B **68**(6), 1085–1094 (1999)
22. Mittleman, D.M.: Twenty years of terahertz imaging. Optics Express **26**(8), 9417–9431 (2018)
23. Ozdemir, A., Polat, K.: Deep learning applications for hyperspectral imaging: a systematic review. Journal of the Institute of Electronics and Computer **2**(1), 39–56 (2020)
24. Popescu, D.C., Ellicar, A.D.: Point spread function estimation for a terahertz imaging system. EURASIP J. Adv. Signal Process. **2010**(1), 575817 (2010)
25. Popescu, D.C., Hellicar, A., Li, Y.: Phantom-based point spread function estimation for terahertz imaging system. In: Proc. Int. Conf. Adv. Concepts for Intell. Vis. Syst. pp. 629–639 (2009)
26. Qin, X., Wang, X., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. In: Proc. of the AAAI Conf. on Artificial Intelligence. pp. 11908–11915 (2020)
27. Recur, B., Younus, A., Salort, S., Mounaix, P., Chassagne, B., Desbarats, P., Caumes, J., Abraham, E.: Investigation on reconstruction methods applied to 3d terahertz computed tomography. Optics Express **19**(6), 5105–5117 (2011)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. Medical Image Comput. Comput.-Assisted Intervention. pp. 234–241 (2015)
29. Rotermund, H.H., Engel, W., Jakubith, S., Von Oertzen, A., Ertl, G.: Methods and application of uv photoelectron microscopy in heterogenous catalysis. Ultramicroscopy **36**(1-3), 164–172 (1991)
30. Round, A.R., Wilkinson, S.J., Hall, C.J., Rogers, K.D., Glatter, O., Wess, T., Ellis, I.O.: A preliminary study of breast cancer diagnosis using laboratory based small angle x-ray scattering. Physics in Medicine & Biology **50**(17), 4159 (2005)
31. Saeedkia, D.: Handbook of terahertz technology for imaging, sensing and communications. Elsevier (2013)
32. Schultz, R., Nielsen, T., Zavaleta, R.J., Wyatt, R., Garner, H.: Hyperspectral imaging: a novel approach for microscopic analysis. Cytometry **43**(4), 239–247 (2001)
33. Slocum, D.M., Slingerland, E.J., Giles, R.H., Goyette, T.M.: Atmospheric absorption of terahertz radiation and water vapor continuum effects. J. Quantitative Spectroscopy and Radiative Transfer **127**, 49–63 (2013)
34. Spies, J.A., Neu, J., Tayvah, U.T., Capobianco, M.D., Pattengale, B., Ostresh, S., Schmuttenmaer, C.A.: Terahertz spectroscopy of emerging materials. The Journal of Physical Chemistry C **124**(41), 22335–22346 (2020)
35. Tuan, T.M., Fujita, H., Dey, N., Ashour, A.S., Ngoc, T.N., Chu, D.T., et al.: Dental diagnosis from x-ray images: an expert system based on fuzzy computing. Biomed. Signal Process. Control **39**, 64–73 (2018)
36. Van Exter, M., Fattinger, C., Grischkowsky, D.: Terahertz time-domain spectroscopy of water vapor. Optics Lett. **14**(20), 1128–1130 (1989)
37. Wong, T.M., Kahl, M., Bolívar, P.H., Kolb, A.: Computational image enhancement for frequency modulated continuous wave (fmcw) thz image. J. Infrared, Millimeter, and Terahertz Waves **40**(7), 775–800 (2019)

38. Xie, X.: A review of recent advances in surface defect detection using texture analysis techniques. ELCVIA: Electron. Lett. Comput. Vis. Iimage Ana. pp. 1–22 (2008)
39. Yujiri, L., Shoucri, M., Moffa, P.: Passive millimeter wave imaging. IEEE Microwave Mag. **4**(3), 39–50 (2003)
40. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: Proc. Int. Conf. Mach. learn. pp. 7354–7363 (2019)
41. Zhang, K., ana Y. Chen, W.Z., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
42. Zhang, K., Zuo, W.M., Zhang, L.: FFDNet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Trans. Image Process. **27**(9), 4608–4622 (2018)
43. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image restoration. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
44. Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, R.B., S.Rosen, M.: Image reconstruction by domain-transform manifold learning. Nature **555**(7697), 487–492 (2018)