

Reconstructing Randomly Masked Spectra Helps DNNs Identify Discriminant Wavenumbers

Yingying Wu, Jinchao Liu, Yan Wang, Stuart Gibson, Margarita Osadchy and Yongchun Fang, *Senior member, IEEE*

Abstract— Nondestructive detection methods, based on vibrational spectroscopy, are vitally important in a wide range of applications including industrial chemistry, pharmacy and national defense. Recently, deep learning has been introduced into vibrational spectroscopy showing great potential. Different from images, text, etc. that offer large labeled data sets, vibrational spectroscopic data is very limited, which requires novel concepts beyond transfer and meta learning. To tackle this, we propose a task-enhanced augmentation network (TeaNet). The key component of TeaNet is a reconstruction module that inputs randomly masked spectra and outputs reconstructed samples that are similar to the original ones, but include additional variations learned from the domain. These augmented samples are used to train the classification model. The reconstruction and prediction parts are trained simultaneously, end-to-end with back-propagation. Results on both synthetic and real-world datasets verified the superiority of the proposed method. In the most difficult synthetic scenarios TeaNet outperformed CNN by 17%. We visualized and analysed the neuron responses of TeaNet and CNN, and found that TeaNet’s ability to identify discriminant wavenumbers was excellent compared to CNN. Our approach is general and can be easily adapted to other domains, offering a solution to more accurate and interpretable few-shot learning.

Index Terms—Masked CNN, deep learning, vibrational spectroscopy

1 INTRODUCTION

VIBRATIONAL spectroscopy, including infrared and Raman, has been widely used for identification and quantification of solids (particles, pellets, powders, films, fibers), liquids (gels, pastes) and gases. Infrared spectroscopy makes use of the vibrational transitions of a molecule, provides fast, non-contact, and non-destructive analysis and has found a wide range of applications in a variety of industries such as petrochemical, chemical, pharmaceutical, cosmetic, food.

An infrared spectrum captures unique molecular information that can be used to characterize a substance and is therefore often referred to as “molecular fingerprint”. Examples of infrared spectra of minerals are shown in Fig. 1. It can be seen that decryption of such spectra is non-trivial, especially in the presence of impurity in the sample and noise caused by the acquisition process. Typically, the analysis of spectra relies heavily upon machine learning methods [1], [2], [3], [4], [5].

Traditionally, automatic recognition of infrared spectra has been based on conventional machine learning methods such as kNN, SVM, PLS [9], [10], [11]. These methods often require pre-processing that typically includes some or all of the following steps: smoothing, baseline correction, multiplicative scatter correction (MSC), standard normal variate

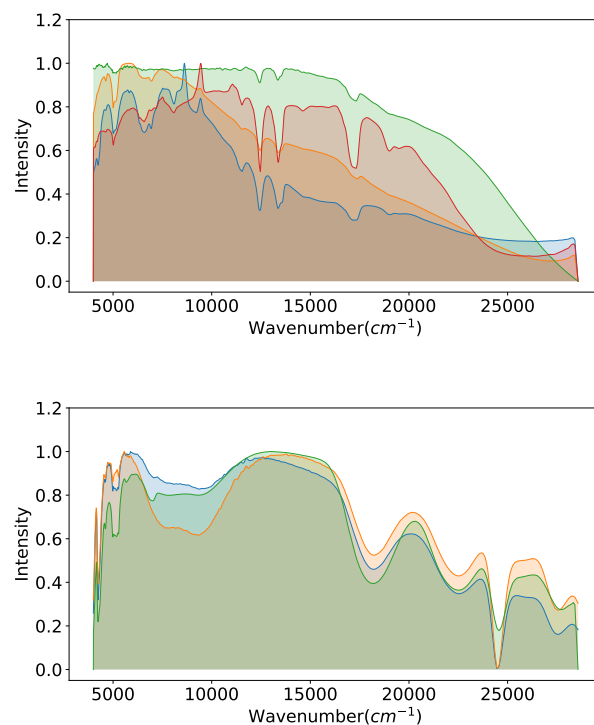


Fig. 1. Examples of near-infrared spectra of two minerals, *Fluorapatite* and *Rhodochrosite*, from the USGS dataset. For each mineral, there are only a handful of samples which differ dramatically from each other. Recognising these minerals from their spectra falls into the category of “few-shot learning” which is rather challenging due to the lack of training samples and the complicated nature of infrared spectra.

- Y. Wu, J. Liu and Y. Fang are with the Institute of Robotics and Automatic Information System (IRAIS), College of Artificial Intelligence, Nankai University, China. e-mail: wuyy@mail.nankai.edu.cn, {liujinchao, fangyc}@nankai.edu.cn (Corresponding author: J. Liu)
- Y. Wang is with VisionMetric Ltd, Canterbury, Kent, UK. e-mail: yan-wangnott@gmail.com
- S. Gibson is with the School of Physics and Astronomy, University of Kent, UK. e-mail: s.j.gibson@kent.ac.uk
- M. Osadchy is with the Department of Computer Science, Haifa University, Israel. e-mail: rita@cs.haifa.ac.il

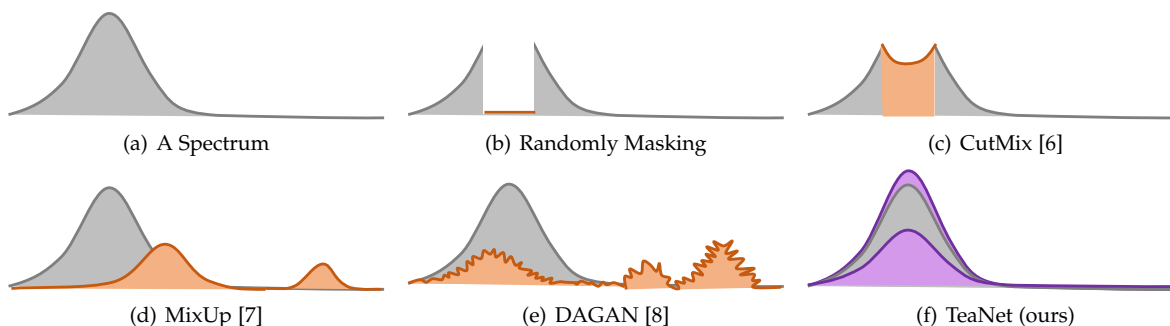


Fig. 2. Illustration of typical augmentation behaviors of existing popular data augmentation strategies and the proposed TeaNet. The variations generated by the existing methods and the proposed TeaNet are indicated in orange and purple respectively. (a) A spectrum with one characteristic peak. (b) Randomly masking generates spectra where a random segment is dropped. (c) CutMix masks the spectrum which is then filled with a random segment from other sample. (d) MixUp combines multiple spectra linearly and conceptually may be regarded as a merging of masking and filling in one step. (e) GAN trained with small data tends to generate augmented sample with no or false variations due to the insufficient training (as observed in our experiments). (f) Our approach TeaNet is able to generate artificial samples with sensible variations.

(SNV) correction, wavenumber selection and principal components analysis (PCA) [12], [13], [14]. Such pre-processing in most cases is non-trivial and could erroneously remove useful features while retaining undesirable noise in the signal.

In recent years, deep neural networks have been proposed for vibrational spectroscopy that have significantly outperformed the conventional machine learning methods in many applications including mineral classification [2], [15], chemical analysis [15], [16], pathogenic bacteria identification [17], rapid detection of COVID-19 causative virus (SARS-CoV-2) [18], [19], [20], metabolite gradients monitoring [21], cytopathology [22], soil properties prediction [15], [23], mine water inrush [24], etc. The use of deep networks in analysing spectra has not only improved the accuracy, but also removed the need for manual, non-trivial, and time consuming preprocessing of spectra.

In many real-world applications of vibrational spectroscopy, there are only a handful of samples/spectra available for each substance. As training deep networks usually requires very large datasets, applications with small data regimes are especially difficult for deep networks. To tackle this problem, previous work proposed and adopted data augmentation techniques [2], [25], employed transfer learning [15], [26] or meta-learning [3]. Despite these efforts, the success of large neural networks in few-shot spectrum recognition has been limited.

Generative adversarial networks (GANs) have shown to be a powerful tool for learning distributions from examples and generating artificial samples. GANs have been used for data augmentation [8], [27], [28] in various tasks. However, we show here that state-of-the-art GANs fail in improving few-shot spectrum recognition (see Table 3). This is expected as training GANs requires even more data than training classifiers. This result motivated us to develop an alternative generative model that trains with self-supervision and makes better use of the small training data (i.e. small volumes of training data).

Besides GANs, there exist a number of popular data augmentation strategies involving masking a signal such as Random Erasing [29] and CutMix [6]. These methods mask random regions of samples which are then “roughly” filled with zeros or random patches from other samples. It has

been shown that they are quite effective when a sufficient number of training samples are available. However, as shown in our experiments, when only a handful of samples are available, these strategies are of little use. When large data is available, introducing a relatively small amount of roughly-filled masked samples could help regularizing the model. In the case of small data where true variations are limited, it is essential to generate artificial samples with sensible variations, instead of near-arbitrary mixups. Fig. 2 shows a graphical illustration of typical variations generated by existing data augmentation strategies and the proposed TeaNet.

Inspired by these augmentation strategies including Random Erasing and CutMix which can be viewed as (random) masking followed by a rough operation of filling, we take this one step further and propose to use a generative model to reconstruct the randomly masked regions to create artificial samples with sensible variations, instead of rough fillings. With the task of “mask-reconstruct”, we can create from a single spectrum many training sample pairs to train the generative model. This is especially useful in small data applications. We named our novel method *Task-Enhanced Augmentation Network* (TeaNet). The mask-reconstruct approach was also inspired by masked language model training BERT [30] and the inpainting task in computer vision [31], [32], [33].

1.1 Contributions

We propose a new method for few-shot classification based on a novel task-enhanced augmentation network. Our approach outperforms previous state-of-the-art methods in infrared spectra classification and it offers interpretability by selecting the most discriminant parts of the input (defined over a range of wavenumbers). Specifically,

- 1) We propose TeaNet, a task-enhanced augmentation network that learns to produce novel samples by filling in the masked part of the input signal with the goal of maximizing the accuracy of the classifier. Source codes of TeaNet are available at <https://github.com/Chaoscendence/TeaNet>.
- 2) Our experiments on synthetic and real-world infrared spectrum datasets showed that TeaNet sig-

nificantly outperforms the state-of-the-art methods in the field.

- 3) We analysed TeaNet and other methods, and showed that TeaNet possessed an excellent ability for discriminating between spectra that were very similar to one another. Our visualization showed that TeaNet was able to locate discriminant regions with much better accuracy than CNN, and therefore offers much better interpretability in addition to superior classification performance.

The main idea behind the TeaNet is general and the architecture can be easily adapted to other domains, offering a general approach to more accurate and interpretable few-shot learning.

1.2 Structure of the Paper

The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 describes the proposed task-enhanced augmentation network (TeaNet). Section 4 and 5 present and discuss the experiments on synthetic data and real-world benchmark datasets. Section 6 summarises our work and discusses potential directions for future research.

2 RELATED WORK

Our work addresses the few-shot learning by augmentation and proposes a generation network which creates synthetic samples that are reconstructed from the real masked samples. It is trained by minimizing a combination of reconstruction and classification losses. We discuss the relevant work in the related areas, specifically, self-supervision through masking and augmentation.

2.1 Data Augmentation

A straightforward solution to the problem of insufficient training data is to synthesize new samples by applying various transforms to the real samples. Simple geometric transforms such as flipping, scaling, rotation and shifting are often applied to image data, whilst patch sampling [34], [35], sample mixing [36], [37] and random erasing and occlusion [29], [38] are among other common strategies. However, all these augmentation methods only involve low-level image operations and do not account for semantic knowledge of the content of the data. We show in our experiments (Section 5) that augmenting training of a deep classification network with spectra obtained by low-level signal operations (we refer to it as CNN-Partial in our experiments) was not always beneficial and even worsened the results in some cases.

Instead of applying low-level transformations, Wang et al. [39] proposed synthesizing new samples at semantic levels by manipulating features. Since features at higher layers of a deep network are characterized by better abstraction, one can assume that introducing some variation at the feature level could offer augmentation with semantic variation.

In recent years, GAN models have shown great capability in synthesizing image data (especially facial images [40])

with high fidelity. GAN models with special architectures have also demonstrated potential for data augmentation [8], [27], [41]. The CWGAN [27] was proposed to augment the electroencephalogram (EEG) data for emotion recognition. It combines WGAN and CGAN and feeds a class label as an additional input to both the generator and discriminator to synthesize samples of specific classes. Antoniou et al. [8] developed DAGAN which consists of a CGAN generator and an improved WGAN critic network. The generator takes a real data item and a random vector as inputs and synthesizes a new data item through an encoder-decoder pipeline. The random vector is sampled from a multi-variate Gaussian distribution to encode the intra-class variations. The generated and the input samples form a combined fake item, while the input sample and another real sample from the same class form a combined real item. The discriminator is trained to distinguish the fake from the real. After training, the original data is then augmented to train a classifier for the target task.

Data augmentation in the feature domain was considered to assist a few-shot learning. Different from our application, [42] makes an assumption on the availability of a large set of labeled data that is related/similar to classes with the small training data. A special case of GAN is trained to model the latent distribution of each novel class given examples from similar “large” classes. The work in [43] uses an auto-encoder to learn differences between same-class pairs of training examples in the latent space, and then applies these “deltas” to small novel data to synthesize samples from the new class. The work in [44] also performs augmentation in the latent space, but the augmentation is guided by the classification loss, which is similar to our method. However, our overall approach is very different from [44]: in our method we create new samples (not features) by reconstructing the masked regions using a reconstruction loss in addition to the classification loss.

2.2 Self-supervised learning

Self-supervision via reconstruction of partially masked inputs has been widely used in language modeling e.g., BERT [30] and more recently in computer vision e.g., ViT [45]. The main goal of these architectures is to utilize the large unlabeled data for pre-training the encoder and then fine-tuning it for the downstream tasks.

A recent work in [46] proposed an asymmetric architecture for masked auto-encoder in which the encoder inputs a small number of randomly selected image patches from an input image with positional information, while the decoder receives the embeddings and the masked symbols and reconstructs the missing parts. It is suggested that masking a very large number of patches in the input encourages the encoder to learn semantic information and not just hole filling. The specific architecture makes the training more efficient. While our approach also employs masking for self-supervision of the reconstruction network, it is used in combination with the classification loss. We do not use the encoder after training, the reconstruction network in our approach is used to enlarge small training data by adding variation to the existing samples via generating the masked parts.

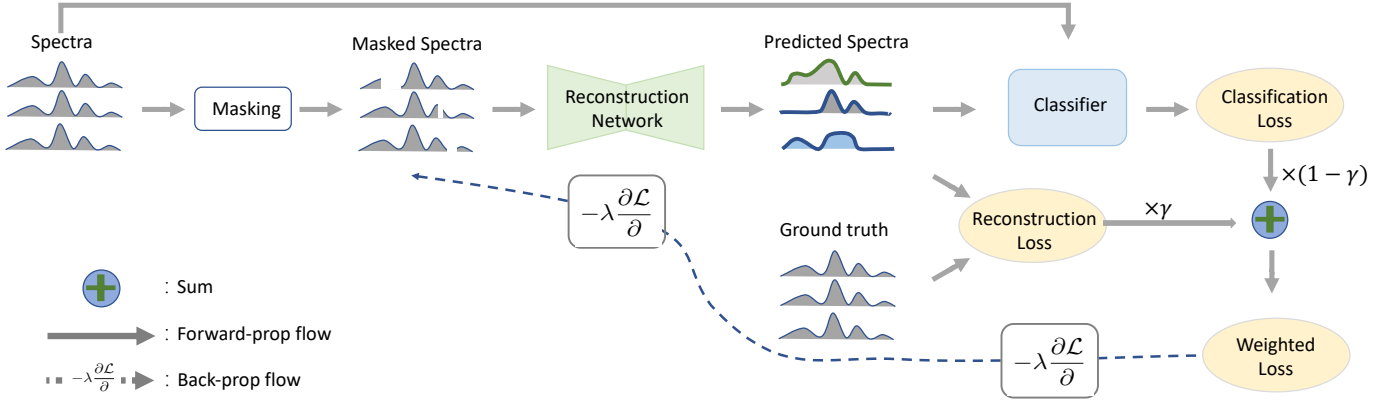


Fig. 3. Diagram of the proposed method TeaNet. It consists of two neural networks, namely reconstruction network and a differential classifier for a given application. The reconstruction network is responsible for generating augmented samples by repairing the masked/corrupt spectra and being led by the classification results. The trainable weights in the reconstruction network are optimised against the weighted sum of the classification and reconstruction losses. While the classifier is trained towards minimizing the classification loss only.

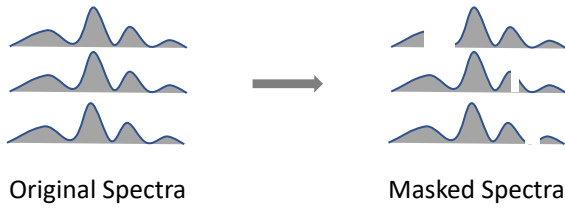


Fig. 4. Examples of masked spectra and the corresponding original ones. Similar to generating pairs to train Siamese networks for few-shot learning [3], generating lots of original-masked spectrum pairs allow us to exploit the training samples in a novel way and thus may help train larger model and achieve better accuracy in the case of few-shot spectrum recognition.

LTSA [47] is a self-supervised feature learning framework where a damage-and-repair network learns to synthesize images with mild artifacts by randomly masking the latent representation of input images, and a discriminator is trained simultaneously to differentiate the images with artifacts from normal ones. The features extracted by the discriminator can later be used for downstream classification tasks. Note that the damage and repair network is relevant to, but different from, our approach. In LTSA, masking or damaging was applied to the latent representation, instead of the input images, in order to generate global artifacts. The success of this scheme greatly relies on good representation to be learned before it can take effect. Our experiments showed that in the case of few-shot learning this approach performed significantly worse than the baseline methods, as shown in Table 3.

3 TEANET: TASK-ENHANCED AUGMENTATION NETWORKS

The small training data problem has been traditionally approached by transfer learning and by meta-learning [48], [49], [50], [51]. We cannot use either of these approaches in the infrared spectrum domain, as transfer learning assumes the existence of a large labeled training set in a similar domain, which does not exist in our application; meta-learning assumes a data set with many episodes from a similar

domain and is limited to a relatively small classification problems (with 5-50 ways). In infrared spectrum recognition the classification problem is usually large, including hundreds of classes.

To tackle the problem of few-shot learning with a large number of classes and no auxiliary training data, we choose data augmentation using generative models. This is a non-trivial task, as training a good generative model with no prior information requires large training sets that we do not have in our setting. This makes regular GANs fail as experimentally confirmed in Table 3. One way to approach this problem is to impose additional constraints on the generative model to reduce its need for a large number of training samples.

Inspired by masking ideas in pre-training large transformers in language modeling [30], and computer vision [45], [46] as well as the inpainting task in images [31], [32], [33], we propose task-enhanced augmentation networks (TeaNet), where a generative model learns through reconstructing samples from randomly masked ones. Note that the masking approach in previous work was used to pretrain the encoder on a very large data and then fine-tune it to the downstream tasks. In this work, we take a different approach and train an auto-encoder not only to reconstruct masked inputs, but to fill the gaps, such that the repaired sample minimizes the classification loss. We augment the original sample with a number of masked samples by randomly masking the original one, as shown in Fig. 4 and repairing them using the auto-encoder. The repaired samples are passed to the classifier and the classification loss on these samples is used to train simultaneously the classifier and the auto-encoder (in addition to the reconstruction loss). A graphical illustration can be found in Fig. 3.

The success of masking in training language models was attributed to the fact that languages are human-generated, and thus are highly semantic and information-dense [46]. Masking words is enough to facilitate the model to capture the semantics of the language. In images, due to spatial redundancy, a missing patch can be recovered from neighboring patches without capturing much of the semantics, thus requiring very heavy masking [46]. The question is

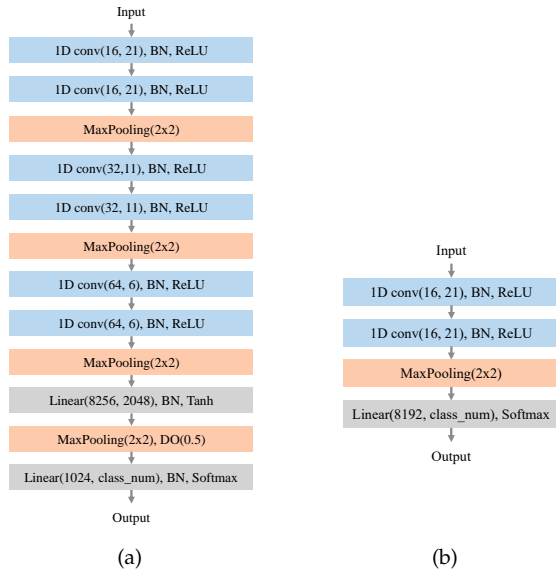


Fig. 5. Network architectures of the classification modules for TeaNet used in the real-world and synthetic datasets. BN stands for batch normalization and DO stands for Dropout.

whether the masking approach would work in spectroscopy and what should be the masking strategy here? The spectroscopy theory in physics shows that any part of an infrared spectrum is related to other parts, which indicates that inferring a missing/masked part from the rest of an infrared spectrum is indeed physically possible, though is very challenging [52].

3.1 Network Architecture

Our network consists of two modules, namely the *reconstruction* module and the *classification* module, which are combined in a pipeline. The classification module is responsible for a spectrum recognition task. Any differentiable model, e.g., convolutional networks, can be used for the classification module. We employ double-pyramid LeNets that have been shown to be very effective in spectrum recognition [2], [3] in our experiments. The detailed architectures especially the number of layers were optimised for different applications (datasets) as shown in Fig. 5.

The reconstruction module creates artificial spectra for domain-specific data augmentation. It inputs a randomly masked real spectrum and outputs a full spectrum, by filling in the missing parts. As the input and the output of this module are both spectra (1D signal), the reconstruction task falls into the category of dense regression where U-Net and its variants are shown to be the most effective architectures [33], [53], [54], [55], [56]. We therefore employ a U-Net architecture in the reconstruction module as shown in Fig. 6. The reconstruction network contains 34 trainable convolutional layers in total and is larger than almost all neural networks previously used for vibrational spectrum recognition.

3.2 Training Protocol

The classification network is trained on real samples along with the corresponding reconstructed samples using the

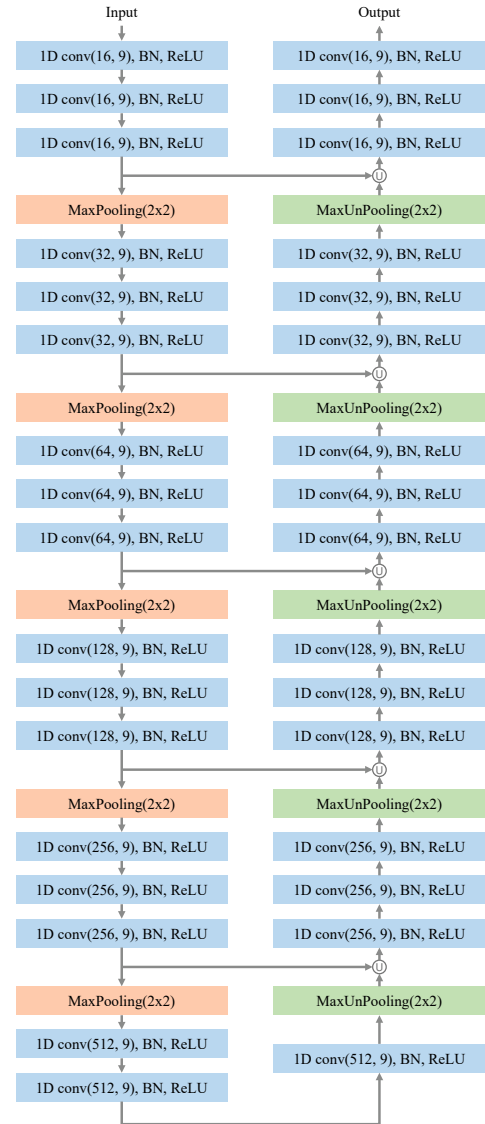


Fig. 6. Network architecture of the reconstruction module in the proposed TeaNet. \oplus stands for feature concatenation.

cross-entropy loss. The reconstruction network is trained to minimize the weighted sum of the classification (cross-entropy) and reconstruction (MSE) losses. The reconstruction network does not input class label of the sample, but the reconstruction is guided by the classification loss, meaning that the reconstructed sample must lie within the class boundaries, but otherwise could be arbitrary. In other words, the classification loss allows the reconstruction to move away from the original sample towards samples that minimize the classification loss. To increase diversity in samples generated from the same input, it is essential to train both networks jointly. If the reconstruction loss dominates the training process of the reconstruction module, it might quickly converge and learn to repair a masked spectrum perfectly. Then, the spectra that it generates would have

no/little difference from the original ones and play no role in data augmentation. The algorithm for training TeaNet is detailed in Algorithm 1.

An important question is how many samples to generate from the original sample. On the one hand, generating lots of original-masked pairs allows us to exploit the training samples in a novel way for training a larger model that may help achieving better accuracy in the case of few-shot spectrum recognition. On the other hand, as the number of pairs grows, the gain from training on those pairs might peak at some point and then start decreasing. We verified this behaviour in our experiments (see Table 5). The extra information/variation embedded in these pairs is bounded, thus many of them will become more and more similar as more masked spectra are generated.

3.3 Masking Schemes

Generation of masks is a key component of the proposed approach. Fig. 4 shows examples of masked spectra and the corresponding original ones. There are several aspects we need to consider in designing the masking scheme. Specifically, should the masks be generated randomly or with fixed position and size and how large they should be?

In general any mask can be represented as a union of small masks $\mathbf{U}\{(a_i, b_i), i = 1, \dots, u\}$ where $a_i \in [0, 1], b_i \in (0, 1]$ denote the relative position of the ends of each segment, and u is the number of segments.

For convenience, we define another variable τ , which is the ratio of mask to spectrum length

$$\tau = \frac{\sum_{i=1}^u |b_i - a_i|}{L} \quad (1)$$

Clearly $\tau \in (0, 1)$. As τ becomes larger, the difficulty of predicting the masked parts of a spectrum increases, requiring the embedding to capture the semantics of the domain and not only trivial dependencies. However, when the mask is too wide, the reconstruction becomes ambiguous which could slow down the learning process.

We consider the following masking schemes controlled by the parameters a and τ :

- MS1: both a and τ are fixed
- MS2: a is fixed, τ is random
- MS3: a is random, τ is fixed
- MS4: both a and τ are random
- MS5: union of several masks of MS4

Our designs of masking schemes include fixed masks (MS1), random masks (MS4 and MS5) and hybrid masks (MS2 and MS3). All schemes, except for MS5, generate continuous masks. MS5 might produce several masked fragments that are disconnected from each other. These designs were inspired by masking schemes in existing work [6], [29], [30] and inpainting tasks in computer vision [31], [32], [33]. In particular, the masking scheme MS4 (random masking) can be regarded as a specific version of Random Erasing (with fixed fillings of zeros) within the spectrum context, except for some specific design choices. Our experiments in Section 5.5.1 showed that reconstructing with MS4 with $a \in \mathcal{U}(0, 1)$ and $\tau \sim \mathcal{U}(0.3, 0.7)$ (were \mathcal{U} stands for uniform distribution) produced the best results.

Algorithm 1: TeaNet: Task-Enhanced Augmentation Network

Input: A training set $S_t = \{(x_i, l_i), i = 1, \dots, N\}$ and a masked set $\{(\bar{x}_i, l_i), i = 1, \dots, N\}$ generated from S_t . A validation set S_v for selecting the best model. The reconstruction and classification modules $\Phi_{\Theta_R}, D_{\Theta_C}$ with trainable weights Θ_R and Θ_C . λ_1, λ_2 are step sizes. $\gamma \in (0, 1)$ is predefined to balance the reconstruction and the classification loss.

for $t = 0$ to T **do**

- (1) The reconstruction module takes a mini-batch of masked samples $\{\bar{x}_i, i = 1, \dots, m\}$ and outputs the repaired ones $\{\hat{x}_i^{(t)}, i = 1, \dots, m\}$, i.e.

$$\hat{x}_i^{(t)} = \Phi_{\Theta_R^{(t)}}(\bar{x}_i)$$

The reconstruction loss $L_R^{(t)}$ is computed as

$$L_R^{(t)} = \frac{1}{m} \sum_{i=1}^m \|\hat{x}_i^{(t)} - x_i\|_2$$

where m is the size of a mini-batch.

- (2) The classification module takes both the original and repaired samples $\{x_i, \hat{x}_i^{(t)}, i = 1, \dots, m\}$, denoted as $\tilde{x}_i^{(t)}$ to simplify the notation, and predicts their labels

$$\tilde{y}_i^{(t)} = D_{\Theta_C^{(t)}}(\tilde{x}_i^{(t)})$$

The cross-entropy loss $L_C^{(t)}$ for classification is computed as

$$L_C^{(t)} = -\frac{1}{m} \sum_{i=1}^m \mathcal{I}(\tilde{y}_i^{(t)} = l_i) \cdot \ln \tilde{y}_i^{(t)}$$

where \mathcal{I} is an indicator function.

- (3) Calculating the joint-loss $L_J^{(t)}$

$$L_J^{(t)} = \gamma * L_R^{(t)} + (1 - \gamma) * L_C^{(t)}$$

and updating Θ_R and Θ_C

$$\Theta_R^{(t+1)} = \Theta_R^{(t)} - \lambda_1 \frac{\partial L_J^{(t)}}{\partial \Theta_R}$$

$$\Theta_C^{(t+1)} = \Theta_C^{(t)} - \lambda_2 \frac{\partial L_C^{(t)}}{\partial \Theta_C}$$

Output: $D_{\Theta_C}^*$ is the classification module with the smallest $L_C^{(t)}$ on the validation set S_v .

4 EXPERIMENT I: ON SYNTHETIC SPECTRA

To analyse the advantage of the proposed TeaNet compared to previous architectures (e.g., CNN [2], [3]), we perform a set of tests on synthetic infrared spectra in which we can control the variation in data, in particular the location and appearance of the peaks (which are the most meaningful parts of a spectrum).

The following experiments showed that TeaNet is able to detect peaks in a wide range of scales. Namely, it is capable of capturing information from both sharp and wide peaks. Conversely, the performance of CNN depended on

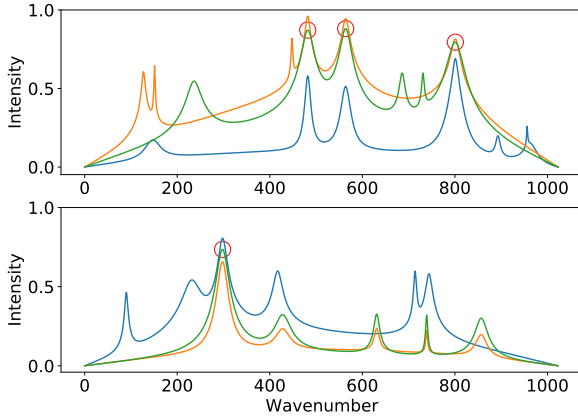


Fig. 7. Synthetic spectra of two classes for few-shot classification created using the Lorentz model. Three samples of each class were plotted here with different colors. The characteristic peaks are marked by circles, while the rest are non-characteristic peaks or simply noise. For these samples, we have set $\vartheta = \frac{10}{3}$ and $\text{FWHM} \in [0, 50]$. Details can be found in section 4.1.2.

the scale of the peaks. It performed well on inputs in which characteristic peaks were sharper and much stronger than non-characteristic peaks. However its performance dropped significantly on spectra with wide and weak peaks.

4.1 Generating synthetic spectra

4.1.1 Physical model

We generate the synthetic spectra using the Lorentz model [57] (in Eq. (2)), which is an infrared spectrum model used in physics:

$$f(p) = C + \sum_{\{\text{FWHM}, A, p_0\}} \frac{A}{1 + \left\{ \frac{p-p_0}{\text{FWHM}/2} \right\}^2} \quad (2)$$

where p denotes a position/wavenumber, p_0 stands for the position of the maximum, FWHM is the full width at half maximum, and A controls the height of a peak. C is a constant, which has been set to be zero in our experiments.

4.1.2 Few-shot Classification Problem

We set up the few-shot classification with N_s ways and M_s shots, namely each classification problem included N_s classes with M_s samples in each. For each class, we first generated K_s^1 characteristic peaks at random. These peaks were shared among all samples in a class. We then produced K_s^2 non-characteristic (noisy) peaks with random or predefined parameters in Eq. (2) and combined them with the characteristic peaks to form samples of the class. We set $N_s = 100$, $M_s = 20$, $K_s^1 \in [1, 3]$, $K_s^2 \in [3, 5]$. The length of each spectrum was 1024. We then generated 20 experiments as explained below with a different set of 100 classes, each class including 20 samples.

4.1.3 Experimental Setting

We denote ϑ to be the ratio of the height of the characteristic and non-characteristic peaks:

$$\vartheta = \frac{A \text{ of a characteristic peak}}{A \text{ of a non-characteristic peak}} \quad (3)$$

ϑ controls the difficulty of the problem, i.e., when ϑ is smaller, it is more difficult for learning models to capture the discriminant information.

To thoroughly investigate the proposed method, we varied two main factors: FWHM (the width) of the characteristic peaks (signals) and ϑ . We tested five typical values of ϑ : 10, 10/3, 1, 0.3, 0.1 and four values of FWHM from $[0, 50]$ to $[150, 200]$. $\text{FWHM} \in [h_l, h_r]$ means when generating the spectra, FWHM were randomly generated from the region $h_l \sim h_r$. Consequently, we performed $5 \times 4 = 20$ sets of experiments in total.

Fig. 7 shows examples of two (out of 100) synthetic classes, generated with $\vartheta = \frac{10}{3}$, $\text{FWHM} \in [0, 50]$. For each class, it shows three samples in different colors. The characteristic peaks are marked by circles, while the rest are non-characteristic peaks. For each the synthetic problem, we randomly partitioned the data set into training, validation and test sets in the ratio of 60%, 20%, 20%.

4.2 Network Architecture and Training Protocol

For comparison, we trained a double-pyramid LeNet, denoted as CNN_Full, for the synthetic classification problems, the detailed architecture of which can be found in Fig. 5(a).

For TeaNet, we have kept the reconstruction module for all the experiments presented in this paper, as shown in Fig. 6. The masking scheme was MS4 with $a \in \mathcal{U}(0, 1)$ and $\tau \sim \mathcal{U}(0.3, 0.7)$. The classification module is the same as the CNN_Full.

To train CNN_Full, we used an Adam optimizer with a learning rate of $1e-4$ and a batch size of 256. We also used Adam optimizers for both reconstruction and classification modules with learning rates of $1e-3$ and $1e-4$ respectively and a batch size of 256. The maximum number of epochs was 300. All the experiments were carried out on multiple NVIDIA GTX-3090(3080) GPUs. These settings have also been adopted for experiments on real-world datasets.

TABLE 1
Performance comparison of TeaNet and CNN on experiments with synthetic spectra.

FWHM \ ϑ	large	medium/small
	small	TeaNet \approx CNN
large	TeaNet \approx CNN	TeaNet \gg CNN

$A \approx B$ - the performance of A and B are similar.

$A \gg B$ - A performs much better than B.

4.3 Results and Analysis

We investigated four typical scenarios shown in Table 1 and the results are shown in Fig. 8 and summarised in Table 1. When ϑ is large, in which case the characteristic peaks dominate the non-characteristic ones, both methods performed equally well. As ϑ decreases, the recognition task become harder. In this setting, TeaNet performed much

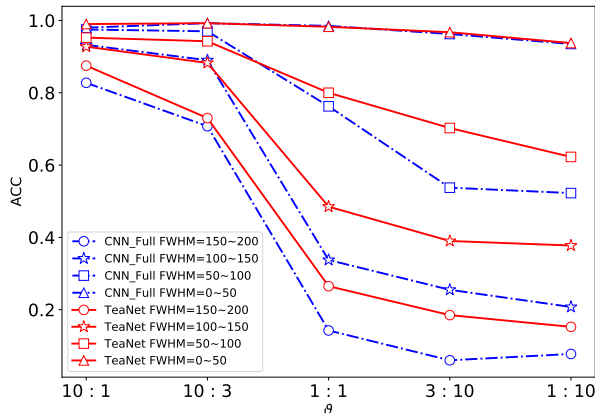


Fig. 8. Comparison of TeaNet with CNN on a synthetic dataset with different settings of characteristic peaks(signals) and non-characteristic peaks(noise or interference). ϑ denotes the ratio of the height of the characteristic peaks and the non-characteristic peaks. Note that when ϑ is medium or small and FWHM is large, the spectrum recognition becomes very difficult. In this case, TeaNet outperformed CNN by a large margin.

better than CNN by a large margin, though the accuracy of both methods dropped.

5 EXPERIMENT II: ON REAL-WORLD SPECTRA

5.1 The Compared Methods

We compared the proposed TeaNet to several alternative augmentation schemes, namely, 1) plain augmentation: CNN_Full [2], 2) inputting partial spectra by randomly masking the full spectra: CNN_Partial, 3) randomly masking followed by filling with random patches from other samples: CutMix [6], 4) randomly combining samples linearly: MixUp [7], 5) generating new samples via GANs: CWGAN [27], DAGAN [8], 6) augmentation in the feature domain: LTSA [47].

CNN_Full was included as the main baseline. We included CNN_Partial for two reasons. First, partial images, aka patches, have been used in computer vision for data augmentation to facilitate the training of large CNNs. Training the classifier on masked spectra is equivalent to training on image patches. The second reason is more hypothetical. Due to randomness, there is a chance that some of the masked spectra would contain only the characteristic peaks. In that ideal case, the masking process would be equivalent to a randomized wavenumber selection, which would lead to a performance improvement. In fact, in real world applications, especially in industry, it is a common practice to use human experts for manually selecting the discriminant wavenumbers and then using them as an input to the conventional machine learning methods, such as PLS. However, in reality, many of the randomly masked spectra would contain lots of non-characteristic peaks, and would require the CNN to detect the characteristic peaks in the masked samples and ignore the noise. So, we may view CNN_Partial as a crude attempt of realizing an automatic wavenumber selection procedure.

CutMix and MixUp are both popular augmentation methods used in computer vision and to some extent both are related to masking for augmentation. In the context of spectrum recognition, CutMix masks a spectrum randomly and then fills the masked region with random patches from other spectra. MixUp combines multiple spectra linearly and conceptually may be regarded as a merging of masking and filling in one step. The rest of the methods (GANs and LTSA) are the natural competitors in data augmentation. Moreover, As CWGAN, DAGAN and LTSA were originally proposed for images, we re-implemented them with necessary modifications associated with 1D signals.

We also tested conventional machine learning methods, such as correlation, kNN, random forest, and support vector machines, as shallow models usually require fewer training examples than the deep models. Moreover, these models are popular among practitioners.

TABLE 2
Benchmark datasets used in this study

Dataset	Type	# Samples	# Classes	# Samples/class
RRUFF_IR	IR	698	430	≈ 1.62
USGS	NIR	887	212	≈ 4.18
ReLab	NIR	6675	855	≈ 7.81

5.2 Datasets and Evaluation Protocol

In our experiments, we used three publicly available datasets, namely RRUFF_IR [58], USGS [59], ReLab. All these datasets contain infrared spectra of minerals. Details can be found in Table 2. It is worth noting that these datasets contain 1.62, 4.18 and 7.81 samples per class on average, which fit the few-shot learning setting.

As the number of samples per class is rather small (see Table 2), we performed the leave-one-out scheme to generate a test set with a single sample per class (chosen at random). The rest of the samples were used for training and validation. We repeated this procedure $N = 10$ times and we reported the averaged accuracy over these experiments.

5.3 Network Architecture and Training Protocol

For the classification module in TeaNet, we adopt a double-pyramid LeNet with the number of convolutional blocks optimised for each benchmark data set. For USGS, the classification module was the same as the one described in 4.2, as shown in Fig. 5(a). For the RRUFF_IR and Relab datasets, the classification module included two convolutional blocks with one dense layer. Details can be found in Fig. 5(b). The architecture of the reconstruction module in TeaNet was the same in all experiments. Please refer to Fig. 6 for more details.

We used Adam optimizer for both reconstruction and classification modules with the learning rate: $1e-3$ and $1e-4$ respectively and 256 batch size. The maximum number of epochs was 300. All the experiments were carried out on multiple NVIDIA GTX-3090(3080) GPUs.

5.4 Results and Analysis

TABLE 3
Classification accuracy (%) of the compared methods on several benchmark spectral datasets

Method \ Dataset	RRUFF_IR	USGS	ReLab
kNN	79.32%±1.56	64.73%±2.31	73.33%±0.87
SVM(linear)	71.22%±1.29	50.70%±2.00	29.91%±1.13
SVM(rbf)	76.33%±1.17	62.17%±2.21	49.52%±1.13
Random Forest	69.46%±3.11	67.44%±1.90	70.68%±0.59
Correlation	74.69%±1.97	38.14%±1.54	51.95%±0.84
LTSA [47]	80.98%±2.60	72.94%±3.44	74.66%±4.04
CWGAN [27]	81.97%±2.13	82.56%±3.31	82.47%±0.99
DAGAN [8]	88.98%±1.21	83.57%±1.42	82.97%±0.92
MixUp [7]	88.85%±1.56	84.58%±1.32	78.44%±0.87
CutMix [6]	87.21%±1.49	85.43%±1.29	82.84%±0.98
CNN_Full [2]	88.03%±0.98	84.81%±1.56	82.06%±1.75
CNN_Partial	86.53%±1.32	86.59%±1.55	83.90%±1.23
TeaNet (ours)	90.07%±1.33	87.37%±1.47	85.43%±1.00

5.4.1 Performance on benchmark datasets

Results in Table 3 show the classification accuracy of the compared methods on the benchmark datasets. First, it is evident that the proposed TeaNet significantly outperforms all the compared methods achieving the state-of-the-art results on all three datasets.

The results of the conventional machine learning methods, such as correlation, kNN, random forest and support vector machines show that even though the conventional methods generally require less data than the deep models, their accuracy is poorer as they are unable to recover useful features and thus require hand-crafted feature engineering.

The accuracy of LTSA on all three datasets is worse than that of the other deep networks by a large margin. In particular, its accuracy is on average nearly 10% lower than the baseline CNN_Full. We believe that even though LTSA utilized inpainting – a concept similar to ours, it was not designed to deal with the few-shot learning scenarios, but rather to learn better representations with the help of pretraining.

We note that CWGAN [27] is merely a plain GAN and serves as a baseline for GAN-like methods in our tests. DAGAN was specifically designed for data augmentation and thus was expected to perform better than the plain CWGAN. Note that on RRUFF_IR and USGS, which have fewer samples per class, DAGAN performed better than CWGAN, yet still worse than the baseline CNN on full spectra. It is not surprising that a plain GAN failed to generate sensible samples when only few samples were provided for its training, as GAN is generally known as a data demanding model. When relatively more samples were available, as in ReLab dataset, both CWGAN and DAGAN performed better and the accuracy increased by about 0.4% and 1.0% respectively.

While similar to approaches for training on image patches, which are beneficial in computer vision problems, CNN_Partial did not perform well in all cases. We observe that, training on partial/masked samples directly improved the accuracy by 1.8% and 1.9% on USGS and ReLab respectively. However, on RRUFF_IR its accuracy dropped dramatically by 1.5%.

Recall that CutMix can be viewed as random masking followed by a naive/rough reconstruction, so if the naive reconstruction helps, one should observe a performance improvement over CNN_Partial. This, however, was not the case. In fact, the accuracy on USGS and ReLab dropped by around 1%. MixUp also experienced a significant performance drop on these two datasets, though it performed relatively well on RRUFF_IR. Overall both methods performed significantly worse than the proposed TeaNet. This may indicate that, in the case of small data where true variations are limited, generating artificial samples with sensible variations is of greater importance, though that dropping/masking a certain amount of regions with a rough reconstruction could help regularize the model when a large amount of data is available.

5.4.2 Analysis of augmented samples

The purpose of the reconstruction module in our approach is to produce dataset augmentation for training the classifier. Thus the best way to evaluate the reconstruction module is by testing whether using reconstructed/augmented samples in classifier training improves the classification accuracy. The results in the previous section show that the augmented samples generated by TeaNet facilitated the training of the classifiers. Moreover, to better understand the behavior of TeaNet, we further investigate the variations (of augmented samples) generated by the reconstruction module.

Visualization of generated variations: To illustrate the variations generated by TeaNet, we plot four augmented spectra for each of two minerals, as shown in Fig. 9. The dissimilarity measured by MSE between these artificial samples and the corresponding original samples were also calculated to be 0.74 ± 0.81 and 3.39 ± 2.02 , the former of which corresponds to the sample shown in the left column of Fig. 9, the latter of which corresponds to the sample in the right column. It can be seen that TeaNet was indeed able to generate sensible variation within each masked region. By randomly masking different wavenumbers/regions, we generated augmented samples where variation occurred in the masked regions. These augmented samples helped TeaNet to identify the discriminant regions of a spectrum.

Explainability of generated variations: It is not unexpected that augmented samples do not have to be human-understandable to assist the training of a complex deep neural network. A good related example is the popular method CutMix which has shown to be effective in image applications, though it generates incoherent artificial samples. On the other hand, we found that TeaNet was indeed able to generate explainable variations. To be specific, these variations generated by TeaNet not only obey the fundamental physical law of continuity, but also reflect physical factors or have performed distinct meanings or functions for learning.

Continuity is a fundamental physical law that a real spectrum should conform to. Though that augmented samples do not necessarily obey physical laws so as to facilitate the training of the classifiers, it is still of interest to investigate whether the augmented samples generated by TeaNet obey the physical law of continuity. It was observed in our

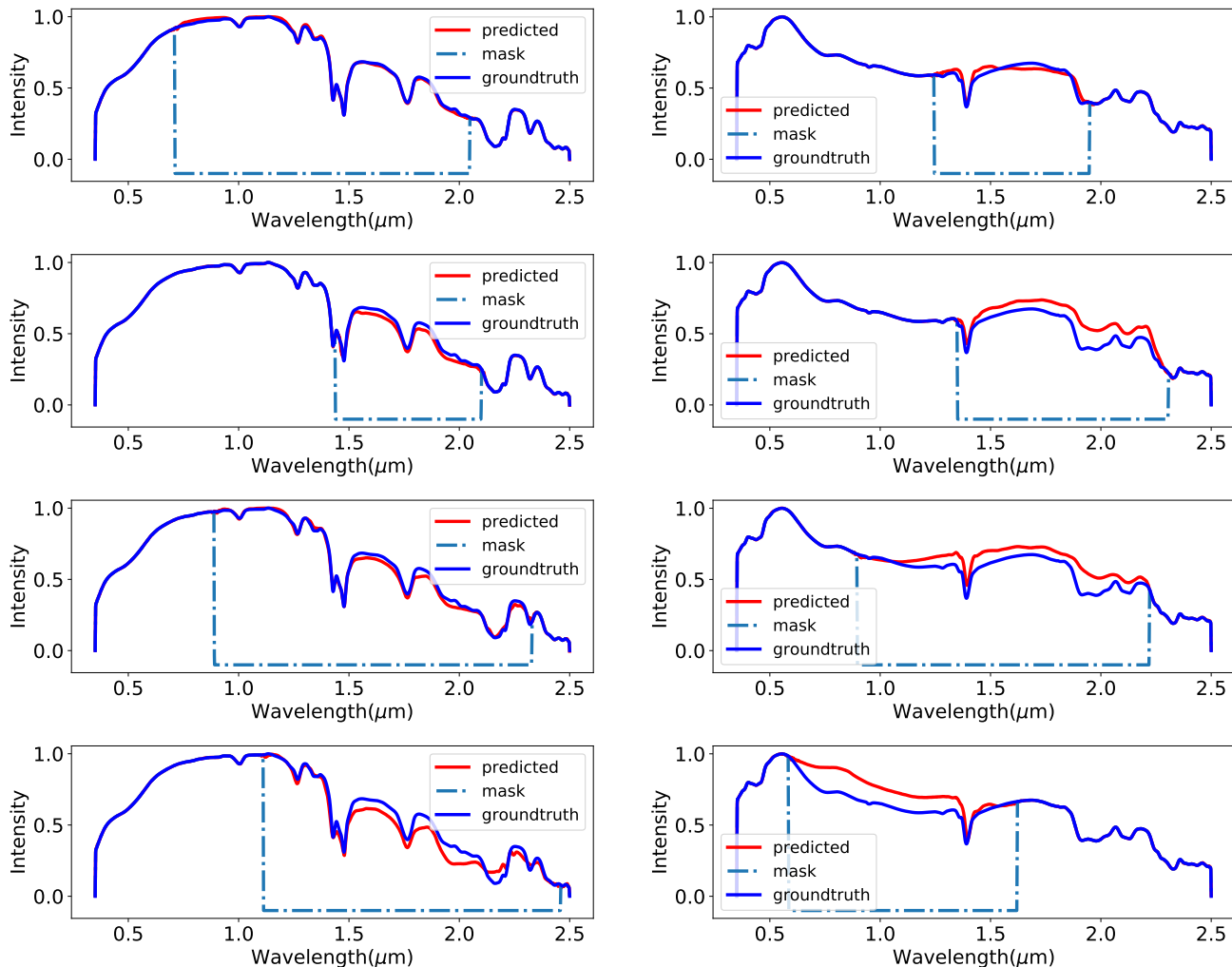


Fig. 9. Variations generated by predicting missing parts of spectra. The original spectra were plotted in blue. The masked regions were indicated by dot lines. The predicted spectra were highlighted in red. The dissimilarity measured by MSE between these artificial samples and the corresponding original samples were calculated as 0.74 ± 0.81 (left) and 3.39 ± 2.02 (right).

experiments that all the generated samples were continuous and the reconstruction module learned to fill the masked region/wavenumbers without breaking the continuity. This suggested that the reconstructed spectra indeed conform to the physical law of continuity. In contrast, CutMix do not preserve the continuity as it simply fills the masked regions with random patches from other samples.

Fig. 10 shows four typical examples of explainable variations generated by TeaNet. Fig. 10(a) demonstrated that TeaNet was able to produce a series of augmented samples by reconstructing different masked regions of the same spectrum. The deformations/variations of these artificial samples bear a resemblance to those induced by physical factors in the real world e.g. temperature changing [60]. Fig. 10(b) showed that TeaNet learned to generate a smooth reconstruction of the masked region of a noisy input. In Fig. 10(c), TeaNet generated the baseline of the input spectrum which implied that it essentially performed a baseline correction which was a popular preprocessing technique for spectrum recognition when non-deep-learning methods were employed [60]. Fig. 10(d) shows an augmented sample where one peak was surgically removed.

Simulated variations which reflect physical factors can be beneficial to the training of the classifiers. Both smoothing and baseline correction have been widely used as preprocessing for vibrational spectroscopy. Masking out a peak “surgically” may be viewed as a smart version of CutMix. All these explainable variations demonstrated that TeaNet was indeed able to generate sensible augmented samples to assist the training of the classifiers.

Cases for which reconstruction failed: We observed that there were a small fraction, around 0.168% on the ReLab dataset, of augmented samples with dramatic variations of large magnitude, the typical cases of which were shown in Fig. 11. Purely from a reconstruction point of view, these might be regarded as failure cases since the augmented samples were vastly different from the original ones. To be particular, the augmented sample shown in the first row of Fig. 11 seems to be arbitrary, but after examination we found that it did not jeopardize the classification. We suspect that this large variation was generated probably because this particular class was easy to classify and therefore could tolerate large variations.

Different from the first one, the variations shown in the

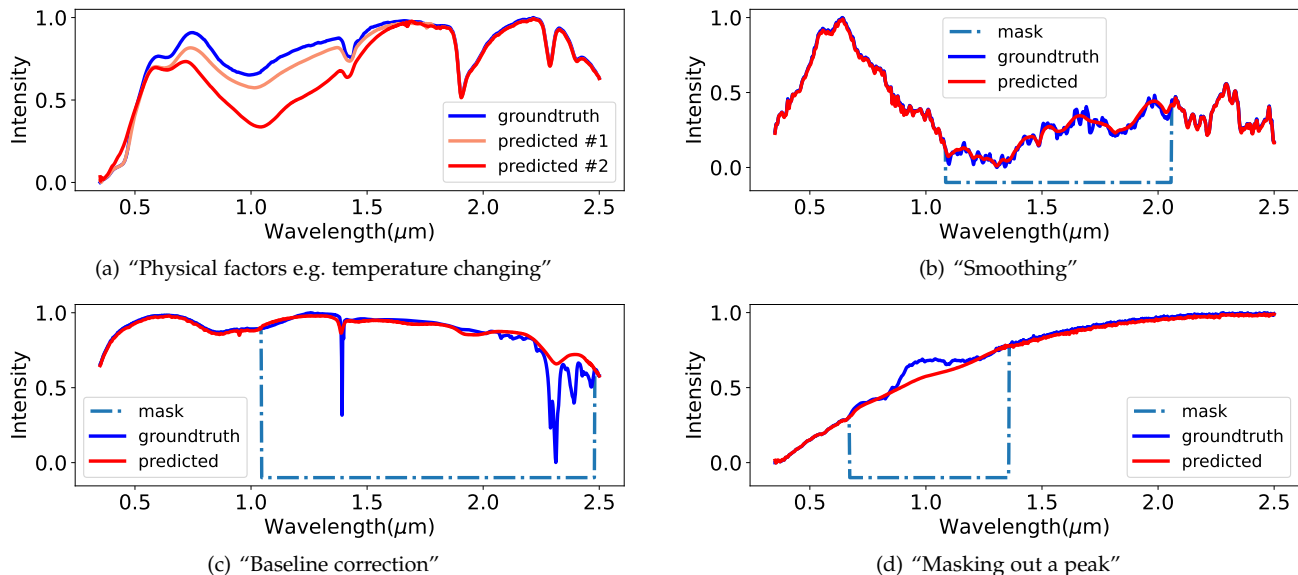


Fig. 10. Typical examples of explainable variations generated by TeaNet. The original spectrum is plotted in blue. The augmented spectra are shown in red and pink. (a) Predicting different masked regions of the same spectrum produced variations which bear a resemblance to those induced by physical factors in the real world e.g. changing temperature [60]. To avoid clutter, the masks are not shown. (b) TeaNet generated a smoothed reconstruction of the masked region of a noisy spectrum. (c) The augmented spectrum corresponds the baseline of the original one on the masked region. Here, TeaNet essentially performed a baseline correction which was a popular preprocessing technique for spectrum recognition when non-deep-learning methods were used [60]. (d) TeaNet generated an augmented sample one peak of which was surgically removed.

second and third rows of Fig. 11 were large, yet sensible, and could still be beneficial, because the corresponding augmented samples could be viewed as the original one with one peak/valley mirrored or trimmed respectively. In this sense, these samples might be meant to be produced to serve our main goal of generating artificial samples with sensible variations to help train the classifier.

The averaged mask size of these samples with large variations was 0.58 ± 0.11 where all the masks were sampled uniformly from 0.3 to 0.7. So we can reduce the number of the failure cases by imposing smaller masks if needed.

5.5 Ablation Analysis

5.5.1 Masking Schemes

Following the protocol outlined in subsection 5.2, we conducted a set of experiments to investigate different masking schemes discussed in Section 3.3. The results are shown in Table 4. For small masks ($\tau < 0.1$), the performance was comparable to the baseline CNN_full, which indicated that variations generated by masking-reconstruction were limited and contributed little to training. With large masks, ($\tau > 0.8$), the accuracy dropped significantly which indicated that the generated variations were inappropriate and contributed negatively to the training.

Overall, the best masking scheme was “MS4” with $\tau \sim \text{URN}(0.3, 0.7)$ with randomized position and length of the mask. Note that MS1 performed significantly worse than the others. This agrees with the fact that some degree of randomization boosts the training of deep models.

5.5.2 Joint vs. separate training of modules in TeaNet

We investigated whether joint training of the reconstruction and classification modules in TeaNet is beneficial over a separate training. We ran the experiments for a varying

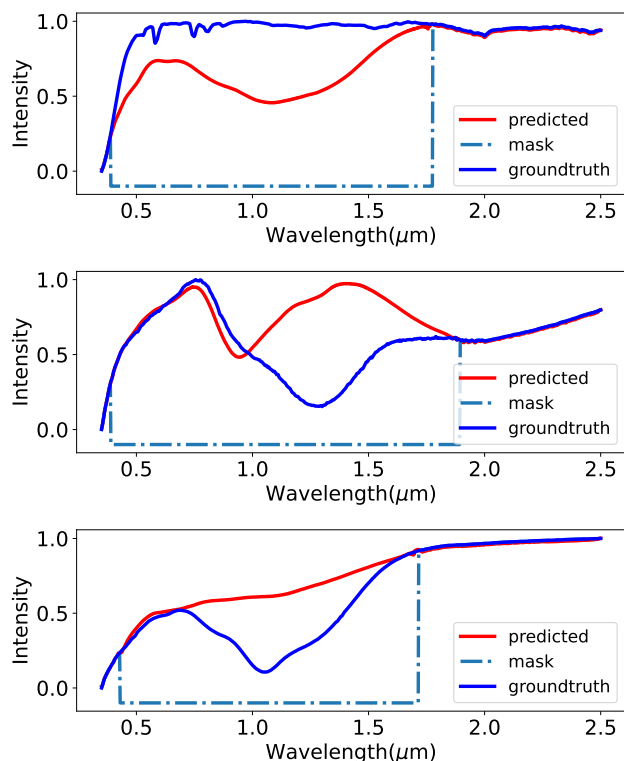


Fig. 11. Large variations generated by TeaNet which might be considered as failure cases from reconstruction point of view as the reconstructed samples were greatly different from the original ones. The original spectra are plotted in blue. The masked regions are indicated by dot lines. The predicted spectra are highlighted in red.

amount of augmentation on all three databases. The results in Table 5 show that joint training of the reconstruction and classification models significantly outperforms training

TABLE 4
Comparison of different masking schemes on the benchmark spectral datasets

Mask \ Dataset	RRUFF_IR	USGS	ReLab
MS1:			
$a = 0, b = 0.5$	87.35%±1.86	82.71%±2.30	80.89%±1.51
$a = 0.5, b = 1$	88.03%±1.53	84.03%±2.00	83.33%±1.26
MS2:			
$a = 0, b \in (0, 1)$	81.97%±3.08	80.85%±2.33	77.93%±1.32
$a = 0.5, b \in (0.5, 1)$	88.78%±1.40	85.35%±2.18	83.96%±0.61
MS3:			
$\tau = 0.01$	87.96%±1.85	86.82%±1.63	82.21%±1.06
$\tau = 0.1$	88.78%±1.96	86.36%±1.39	83.04%±1.02
$\tau = 0.2$	88.64%±1.14	86.67%±1.66	84.28%±0.79
$\tau = 0.5$	90.07%±1.81	86.98%±1.73	84.89%±0.95
$\tau = 0.8$	87.85%±2.13	85.58%±1.90	83.70%±0.91
$\tau = 0.9$	87.08%±1.83	83.26%±2.03	81.55%±1.19
$\tau = 0.99$	72.18%±2.36	73.41%±2.69	70.12%±1.60
MS4:			
$\tau \sim \mathcal{U}(0.1, 0.9)$	89.32%±1.75	84.73%±2.08	84.74%±0.79
$\tau \sim \mathcal{U}(0.3, 0.7)$	90.07%±1.33	87.37%±1.47	85.43%±1.00
MS5:			
$\tau_i \sim \mathcal{U}(0.1, 0.3)$			
$\sum_i \tau_i \sim \mathcal{U}(0.3, 0.7)$	89.53%±0.87	86.13%±2.15	85.45%±0.62

\mathcal{U} stands for uniform random number sampler.
 a, b are the two ends of a mask. τ is defined in Equation (1).

TABLE 5
Comparison of joint vs. separate training of two modules in TeaNet. ($\times n$) corresponds to the number of masked samples generated for training the network

Method \ Dataset	RRUFF_IR	USGS	ReLab
Separately($\times 5$)			
Jointly($\times 5$)	86.94%±2.06	84.81%±1.56	83.57%±1.21
Separately($\times 10$)			
Jointly($\times 10$)	89.05%±1.03	86.12%±1.85	84.69%±0.50
Separately($\times 20$)			
Jointly($\times 20$)	87.96%±0.92	84.65%±1.38	83.05%±1.22
Separately($\times 20$)			
Jointly($\times 20$)	90.07%±1.33	87.37%±1.47	85.43%±1.00
Separately($\times 20$)			
Jointly($\times 20$)	87.76%±1.83	84.26%±1.67	83.78%±1.42
Separately($\times 20$)			
Jointly($\times 20$)	89.66%±1.00	87.06%±1.55	85.48%±1.18

them separately.

5.5.3 Masking vs. Reconstruction

In our approach, both components of masking and reconstruction are crucial and together they cooperate to generate artificial samples with sensible variations. Without masking, one would need to generate artificial samples from random noise which demands a large amount of data to train. As shown in Table 3, GANs were even inferior to the baseline CNN_Full. Without reconstruction, as showed in Table 3, only masking e.g. CNN_Partial or masking with rough reconstruction e.g. CutMix and MixUp, they all performed poorly. Masking followed by careful sensible reconstruction as in the proposed TeaNet indeed led to a significant improvement.

5.6 Model Complexity and Computational Cost

In this section, we discuss the model complexity and the computational cost for the proposed methods with a comparison to existing generative model-based methods. As all

TABLE 6
Model complexity and the computational cost for the training of the proposed TeaNet compared to existing methods, excluding the classification model. For inference, only the classification model is to be used.

Method	Model Size	GPU Memo.	Forw. Pass	Back. Pass
LTSA [47]	280.23 M	5.12 G	112.17 ms	115.94 ms
CWGAN [27]	17.64 M	1.81 G	4.39 ms	7.96 ms
DAGAN [8]	1.85 M	1.97 G	33.95 ms	101.24 ms
TeaNet (ours)	43.36 M	2.06 G	17.27 ms	40.64 ms

the compared methods are proposed for data augmentation, and play no direct role in inference where only the classification is to be used, we calculated related metrics only for the training. Results can be found in Table 6. The required GPU memory and run-time were computed within an epoch with a batch size of 16 on a server with a NVIDIA GTX 3090 GPU and Intel(R) Xeon(R) CPU E5-2678. It can be seen that the proposed TeaNet used comparable GPU memory to DAGAN and CWGAN, ran much faster than DAGAN, and yet was slower than CWGAN. We conclude that with the augmented samples, generated by reconstructing from randomly masked spectra, we were able to train a relatively large model and achieved superior performance with computational costs that are comparable to the GANs.

5.7 Interpreting TeaNet on Challenging Cases

Our experiments on real and synthetic data showed a clear advantage of TeaNet over other methods. We subsequently tried to interpret how the TeaNet makes decisions in comparison to the baseline CNN. To this end, we focus on the most challenging cases of very similar samples that belong to different classes. Fig. 12 depicts two such examples that were correctly classified by TeaNet, but confused the CNN. Namely, the first and second rows in Fig. 12 show samples in the training and test sets respectively. The last row shows samples from the training set of the class misclassified by CNN. It can be seen that in both cases, the similarity between the spectra of the correct and the misclassified classes is very high, and therefore it is challenging to distinguish between them. In fact, visually it is almost impossible for non-experts to make correct classification.

By analysing these cases, we demonstrate a limited ability of CNN in identifying discriminant wavenumbers especially when dealing with very similar spectra. In contrast, TeaNet showed superior performance in locating discriminant regions and achieving high classification accuracy as we show below.

To interpret how the networks make decisions, we visualize their neuron responses to spectra using Grad-CAM [63], and quantitatively evaluate what discriminant information was extracted and utilized for classification. To this end we define two measures; the *similarity of two responses* \mathcal{S} and the *attention on discriminant wavenumbers* \mathcal{A} .

$$\mathcal{S} = \frac{\mathbf{v}_r \cdot \mathbf{v}_s}{\|\mathbf{v}_r\| \|\mathbf{v}_s\|} \quad (4)$$

where \mathbf{v}_r and \mathbf{v}_s are two responses/vectors which correspond to a segment or a full spectrum. \mathcal{S} measures how similar two responses are. For instance, $\mathcal{S} = 1$ means the

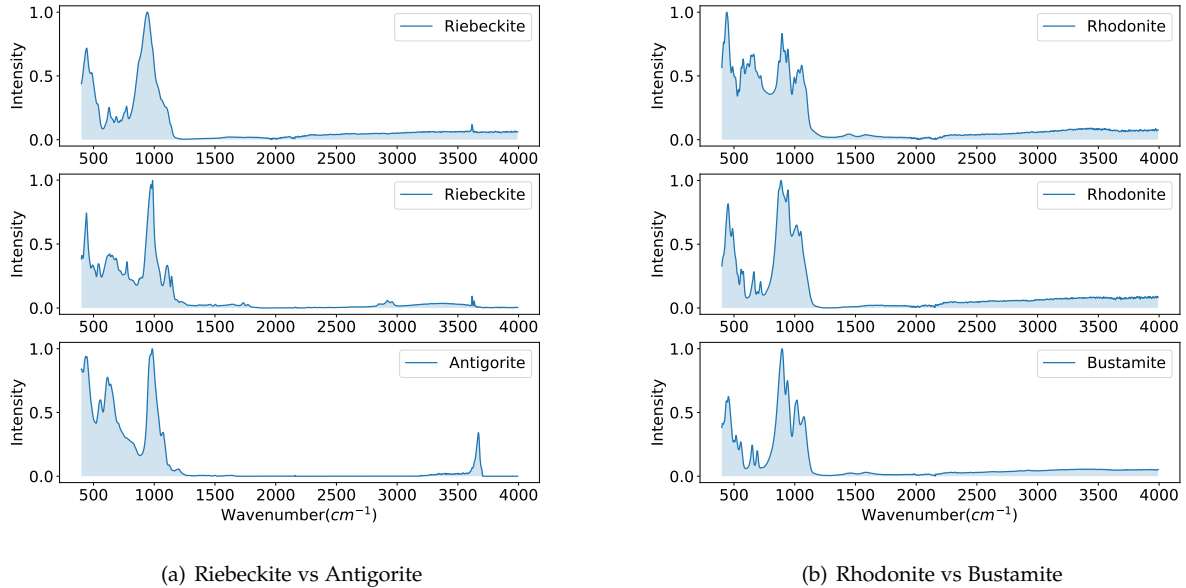


Fig. 12. Challenging cases where TeaNet succeeded while the state-of-the-art CNN failed. Each figure presents a case where the first and second rows plot samples in the training and test sets, respectively. The last row plots samples from the training set of the class misclassified by CNN. It can be seen that visually they are nearly indistinguishable.

TABLE 7
Analysis of neuron responses of TeaNet and CNN on different regions of IR spectra of Riebeckite and Antigorite.

	R_1	R_2	R_3	R_4	R_{1-4}
Wavenumber (cm^{-1})	400 – 475	476 – 600	601 – 1200	3600 – 3700	-
Assigned Bands [61]	T($\text{Fe}^{2+} - \text{OH}$) T($\text{Mg}^{2+} - \text{OH}$)	bending Si-O-Si	$v_s(\text{Si}-\text{O}_b-\text{Si})$ $v_{as}(\text{Si}-\text{O}_b-\text{Si})$ $v_{as}(\text{O}-\text{Si}-\text{O})$	OH stretching	-
S of CNN	0.795	0.567	0.809	0.398	0.748 ↓
\mathcal{A} of CNN (<i>Riebeckite</i>)	0.066	0.054	0.446	0.010	0.576 ↑
\mathcal{A} of CNN (<i>Antigorite</i>)	0.082	0.094	0.355	0.055	0.586 ↑
S of TeaNet	0.587	0.572	0.382	0.132	0.331 ↓
\mathcal{A} of TeaNet (<i>Riebeckite</i>)	0.278	0.117	0.505	0.019	0.919 ↑
\mathcal{A} of TeaNet (<i>Antigorite</i>)	0.086	0.130	0.611	0.136	0.963 ↑

T - Transational modes; v_s - Symmetric stretching; v_{as} - Asymmetric stretching; s_{nr} - Similarity of Neuron Responses; l_{nr} - Intensity of Neuron Responses; ↓ - lower, better; ↑ - higher, better.

corresponding network reacts almost constantly to any spectra, which of course may lead to a very low classification accuracy.

$$\mathcal{A}_j = \frac{\|\mathbf{v}_j\|_1}{\|\mathbf{v}\|_1} \quad (5)$$

where \mathbf{v}_j is the j^{th} segment of a spectrum \mathbf{v} , and $\mathcal{A}_j \in (0, 1]$ is the normalized sum of the intensity of a response on the j^{th} segment (region). \mathcal{A}_j on discriminant regions/segments indicates how the intensity/energy is distributed over the spectrum. Ideally, high values of \mathcal{A} should correspond to the discriminative segments in the input, which in turn should correlate with improved classification accuracy.

Due to space limitations, we only present detailed analysis for two challenging examples shown in Fig. 12. Addi-

tional examples of challenging cases are shown in supplementary information.

5.7.1 Antigorite vs Riebeckite

Fig. 12(a) shows the spectra of the minerals *Riebeckite* and *Antigorite*, the chemical formula of which are $\text{Na}_2(\text{Fe}_3^{2+}\text{Fe}_2^{3+})\text{Si}_8\text{O}_{22}(\text{OH})_2$ and $\text{Mg}_3\text{Si}_2\text{O}_5(\text{OH})_4$ respectively. According to [61], we identified all the discriminant regions (wavenumbers) on the spectra of *Riebeckite*. We categorized all the characteristic peaks into four regions $\{R_i, i = 1, \dots, 4\}$ and marked the peaks as R_i^k where k is the index of the peak within its region.

Fig. 13 shows the discriminant peaks on the spectrum and the corresponding matches (peaks) on the neuron responses of TeaNet and CNN, which were found as the closest peak within a small neighborhood, 0.5% of the

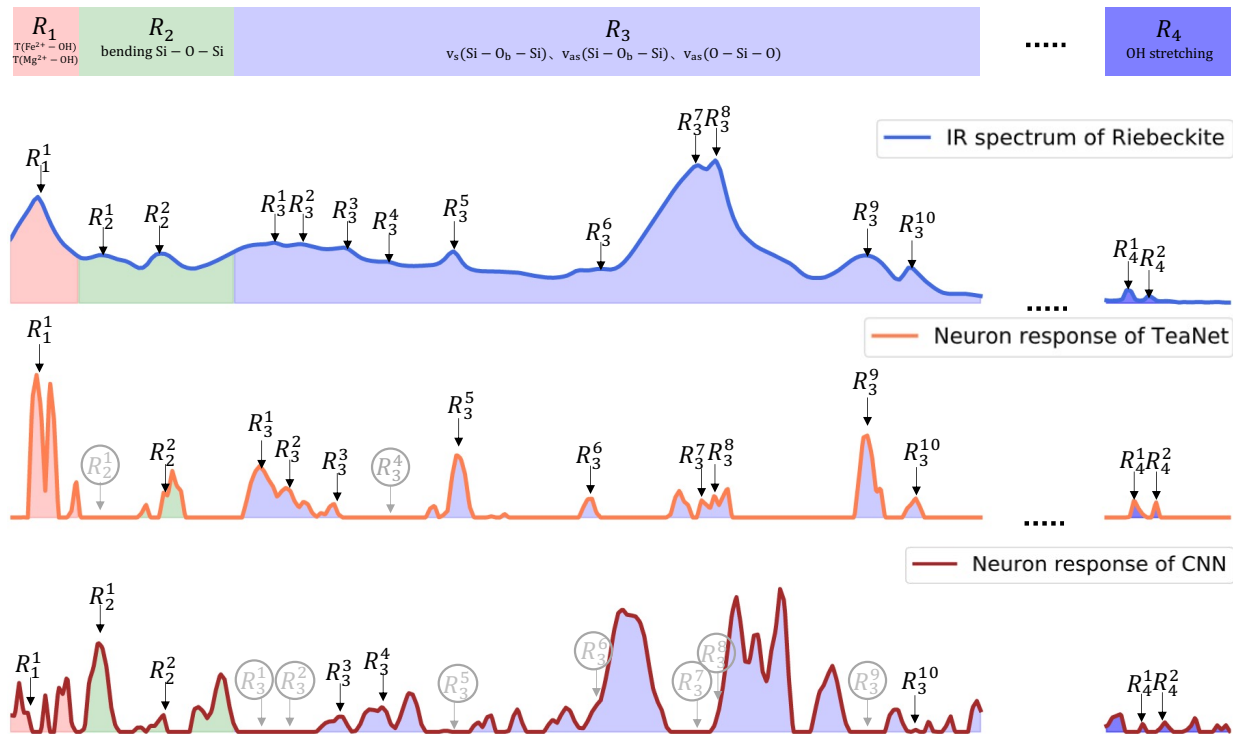


Fig. 13. Comparison of neuron responses of TeaNet and CNN on a challenging case of discriminating between “Riebeckite” and “Antigorite”, the spectra of which, see Fig. 12, is almost indistinguishable visually from one another. On this case, TeaNet succeeded while the CNN failed. According to [61], we categorized the peaks into $\{R_i, i = 1, \dots, 4\}$ and marked the peaks as R_i^k where k is the index of the peak within its region for convenience. We visualize the neuron responses(activation maps) of TeaNet and CNN by Grad-CAM, and highlighted the matches to the peaks of the spectrum. The peaks which were ignored by either method were circled and marked in grey. It can be seen that responses of TeaNet to almost all of the characteristic peaks are accurate and well-shaped, while CNN missed quite a few.

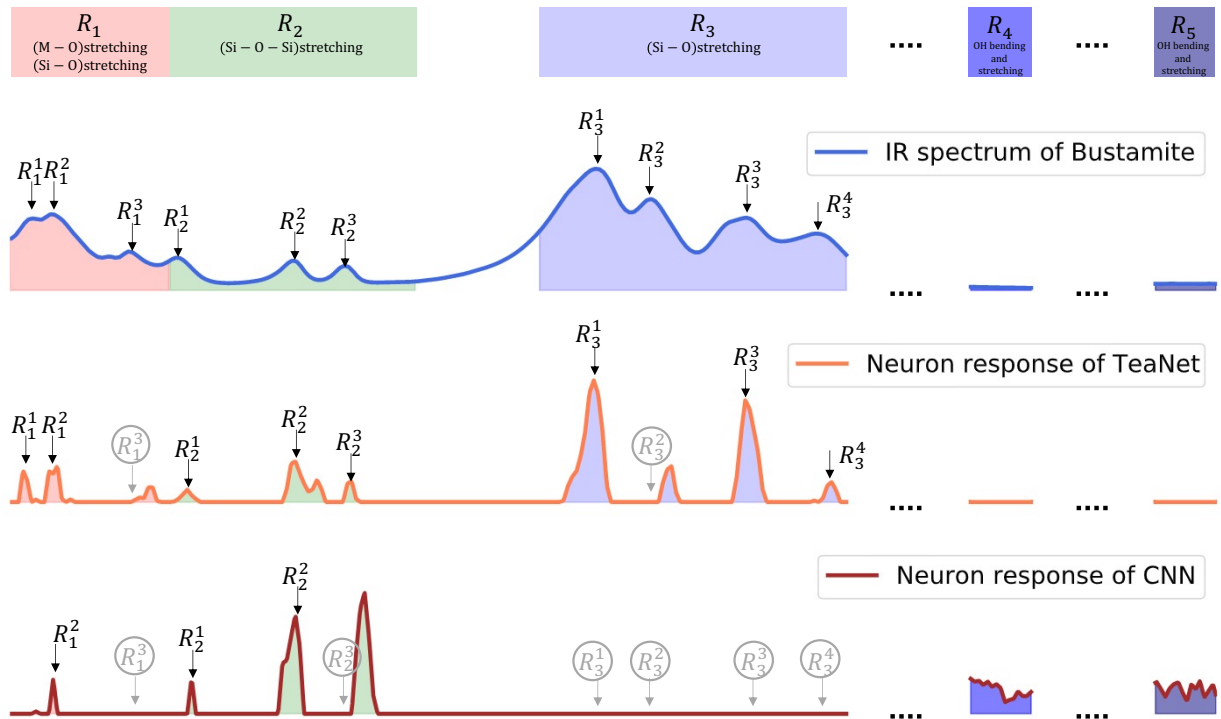


Fig. 14. Comparison of neuron responses of TeaNet and CNN on a challenging case of discriminating between “Riebeckite” and “Antigorite”, the spectra of which, see Fig. 12, is almost indistinguishable visually from one another. On this case, TeaNet succeeded while the state-of-the-art CNN failed. According to [62], we categorized the peaks into $\{R_i, i = 1, \dots, 5\}$ and marked the peaks as R_i^k where k is the index of the peak within its region for convenience. We visualize the neuron responses(activation maps) of TeaNet and CNN by Grad-CAM, and highlighted the matches to the peaks of the spectrum. The peaks which were ignored by either method were circled and marked in grey. It can be seen that responses of TeaNet to almost all of the characteristic peaks are accurate and well-shaped, while CNN missed quite a few.

TABLE 8
Analysis of neuron responses of TeaNet and CNN on different regions of IR spectra of Rhodonite and Bustamite.

	R_1	R_2	R_3	R_4	R_5	R_{1-5}
Wavenumber (cm^{-1})	400 – 550	551 – 750	850 – 1100	1600 – 1650	3400 – 3450	-
Assigned Bands [62]	(M–O) stretching (Si–O) bending	(Si–O–Si) stretching	(Si–O) stretching	OH bending and stretching		-
\mathcal{S} of CNN	0.090	0.081	0.000	0.803	0.798	0.347 ↓
\mathcal{A} of CNN (<i>Rhodonite</i>)	0.070	0.071	0.114	0.009	0.016	0.281 ↑
\mathcal{A} of CNN (<i>Bustamite</i>)	0.008	0.079	0.000	0.028	0.038	0.153 ↑
\mathcal{S} of TeaNet	0.325	0.054	0.533	0.000	0.000	0.392 ↓
\mathcal{A} of TeaNet (<i>Rhodonite</i>)	0.231	0.139	0.370	0.000	0.003	0.743 ↑
\mathcal{A} of TeaNet (<i>Bustamite</i>)	0.136	0.145	0.710	0.000	0.000	0.992 ↑

s_{nr} - Similarity of Neuron Responses; l_{nr} - Intensity of Neuron Responses; ↓ - lower, better; ↑ - higher, better.

spectrum length, of the discriminant peaks on the original spectrum. Peaks which were ignored by either method were circled and marked in grey. It can be seen that responses of TeaNet to almost all of the characteristic peaks are accurate and well-shaped, while CNN missed 7 out of 15 peaks. The ability to capture more discriminant peaks enables TeaNet to distinguish spectra with high similarity.

Further more, to quantitatively analyse the neuron responses, we calculated \mathcal{S} and \mathcal{A} on each region of $\{R_i, i = 1, \dots, 4\}$ individually as well as the combination of these regions. Results are shown in Table 7. It can be seen that \mathcal{S} of TeaNet on full spectra was 0.331 which is significantly lower than 0.748 of CNN. This indicates that compared to CNN, TeaNet reacted differently to spectra of different minerals which enables it to distinguish spectra with minor differences. \mathcal{A} of CNN on the regions combined was less than 60%, while that of TeaNet was more than 90%. This demonstrated that compared to CNN, TeaNet could better identify the discriminant regions, make predictions based upon those selected wavenumbers and therefore achieve higher accuracy.

5.7.2 Bustamite vs Rhodonite

Fig. 12(b) shows the spectra of the minerals, *Bustamite* and *Rhodonite*, the chemical formula of which are $\text{Mn}_2\text{Ca}_2\text{MnCa}(\text{Si}_3\text{O}_9)_2$ and $\text{CaMn}_3\text{Mn}(\text{Si}_5\text{O}_{15})$ respectively.

Similar to Fig. 13, we plot discriminant peaks on the original spectrum and their matches on neuron responses in Fig. 14. Evidently TeaNet captured much more discriminant peaks than CNN.

We calculated \mathcal{S} and \mathcal{A} on full spectra as well as 5 discriminant regions [62]. Results are presented in Table 8. It can be seen that in this case \mathcal{S} of TeaNet and CNN are 0.392 and 0.347 respectively, which indicates both reacted differently to the compared minerals. On the other hand, \mathcal{A} of CNN to both minerals are 28.1% and 15.3% respectively, which is extremely low and indicates that CNN actually failed to concentrate on the discriminant regions. Indeed, by checking the neuron responses, we observed that when fed with Bustamite, CNN completely missed the discriminant region R_3 .

This confirms that CNN’s ability to locate the discriminant regions is limited, especially for challenging cases as

the above-mentioned ones. While enhanced by the task of reconstruction of masked spectra, even with the same network architecture, TeaNet showed great potential in identifying discriminant wavenumbers and achieved superior performance of classification. In this sense, TeaNet seems to be able to implicitly perform wavenumbers selection, a technique which has been widely used in industrial applications of vibrational spectroscopy.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a task-enhanced augmentation network for a few-shot vibrational spectrum recognition. By reconstructing randomly masked samples, the network is encouraged to generate augmented samples with meaningful variation in the masked regions. Results on both synthetic and real-world datasets verified the superiority of the proposed method. Moreover, by visualizing and analysing the neuron responses of TeaNet, we demonstrated TeaNet’s excellent ability of identifying discriminant wavenumbers, which allows it to distinguish spectra with high similarity in challenging cases where the state-of-the-art methods including CNN tend to fail. Future research will investigate additional losses which measure the diversity of artificial samples and learnable masking parameters to further improve the diversity and efficiency of the sample-generating process, and ultimately increase the classification accuracy.

ACKNOWLEDGMENT

This research was supported by National Natural Science Foundation of China No. 62076140.

REFERENCES

- [1] E. Bouveresse and D. Massart, “Improvement of the piecewise direct standardisation procedure for the transfer of nir spectra for multivariate calibration,” *Chemom. Intell. Lab. Syst.*, vol. 32, no. 2, pp. 201–213, 1996.
- [2] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, and S. J. Gibson, “Deep convolutional neural networks for raman spectrum recognition: a unified solution,” *Analyst*, vol. 142, no. 21, pp. 4067–4074, 2017.
- [3] J. Liu, S. J. Gibson, J. Mills, and M. Osadchy, “Dynamic spectrum matching with one-shot learning,” *Chemom. Intell. Lab. Syst.*, vol. 184, pp. 175–181, 2019.

- [4] C. Cui and T. Fearn, "Modern practical convolutional neural networks for multivariate regression: Applications to nir calibration," *Chemom. Intell. Lab. Syst.*, vol. 182, pp. 9–20, 2018.
- [5] X. Cui, Y. Sun, W. Cai, and X. Shao, "Chemometric methods for extracting information from temperature-dependent near-infrared spectra," *Sci. China Chem.*, vol. 62, no. 5, pp. 583–591, 2019.
- [6] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.
- [7] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Int. Conf. Learn. Representations*, 2018.
- [8] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.
- [9] Q. Zheng, J. Li, L. Yang, B. Zheng, J. Wang, N. Lv, J. Luo, F. L. Martin, D. Liu, and J. He, "Raman spectroscopy as a potential diagnostic tool to analyse biochemical alterations in lung cancer," *Analyst*, vol. 145, no. 2, pp. 385–392, 2020.
- [10] M. Sattler, C. Bessant, J. Smith, and N. Stone, "Investigation of support vector machines and raman spectroscopy for lymph node diagnostics," *Analyst*, vol. 135, no. 5, pp. 895–901, 2010.
- [11] L. C. Lee, C.-Y. Liong, and A. A. Jemain, "Partial least squares-discriminant analysis (pls-da) for classification of high-dimensional (hd) data: a review of contemporary practice strategies and knowledge gaps," *Analyst*, vol. 143, no. 15, pp. 3526–3539, 2018.
- [12] Z.-M. Zhang, S. Chen, and Y.-Z. Liang, "Baseline correction using adaptive iteratively reweighted penalized least squares," *Analyst*, vol. 135, no. 5, pp. 1138–1146, 2010.
- [13] J. Peng, S. Peng, Q. Xie, and J. Wei, "Baseline correction combined partial least squares algorithm and its application in on-line fourier transform infrared quantitative analysis," *Anal. Chim. Acta*, vol. 690, no. 2, pp. 162–168, 2011.
- [14] R. Helin, U. G. Indahl, O. Tomic, and K. H. Liland, "On the possible benefits of deep learning for spectral preprocessing," *J. Chemom.*, vol. 36, no. 2, p. e3374, 2022.
- [15] L. Liu, M. Ji, and M. Buchroithner, "Transfer learning for soil spectroscopy based on convolutional neural networks and its application in soil clay content mapping using hyperspectral imagery," *Sensors*, vol. 18, no. 9, p. 3169, 2018.
- [16] X. Fan, W. Ming, H. Zeng, Z. Zhang, and H. Lu, "Deep learning-based component identification for the raman spectra of mixtures," *Analyst*, vol. 144, no. 5, pp. 1789–1798, 2019.
- [17] C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. Saleh, S. Ermon, and J. Dionne, "Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning," *Nat. Commun.*, vol. 10, no. 1, pp. 1–8, 2019.
- [18] C. Carlomagno, D. Bertazioli, A. Gualerzi, S. Picciolini, P. Banfi, A. Lax, E. Messina, J. Navarro, L. Bianchi, A. Caronni *et al.*, "Covid-19 salivary raman fingerprint: innovative approach for the detection of current and past sars-cov-2 infections," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021.
- [19] D. Zhang, X. Zhang, R. Ma, S. Deng, X. Wang, X. Wang, X. Zhang, X. Huang, Y. Liu, G. Li *et al.*, "Ultra-fast and onsite interrogation of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) in waters via surface enhanced raman scattering (sers)," *Water Res.*, vol. 200, p. 117243, 2021.
- [20] K. Ember, F. Daoust, M. Mahfoud, F. Dallaire, E. Z. Ahmad, T. Tran, A. Plante, M.-K. Diop, T. Nguyen, A. St-Georges-Robillard *et al.*, "Saliva-based detection of covid-19 infection in a real-world setting using reagent-free raman spectroscopy and machine learning," *J. Biomed. Opt.*, vol. 27, no. 2, p. 025002, 2022.
- [21] F. Lussier, D. Missirlis, J. P. Spatz, and J.-F. Masson, "Machine-learning-driven surface-enhanced raman scattering optophysiology reveals multiplexed metabolite gradients near cells," *ACS nano*, vol. 13, no. 2, pp. 1403–1411, 2019.
- [22] S. D. Krauß, R. Roy, H. K. Yosef, T. Lehtonen, S. F. El-Mashtoly, K. Gerwert, and A. Mosig, "Hierarchical deep convolutional neural networks combine spectral and spatial information for highly accurate raman-microscopy-based cytopathology," *J. Biophotonics*, vol. 11, no. 10, p. e201800022, 2018.
- [23] J. Padarian, B. Minasny, and A. McBratney, "Using deep learning to predict soil properties from regional spectral data," *Geoderma Regional*, vol. 16, p. e00198, 2019.
- [24] F. Hu, M. Zhou, P. Yan, D. Li, W. Lai, K. Bian, and R. Dai, "Identification of mine water intrusion using laser-induced fluorescence spectroscopy combined with one-dimensional convolutional neural network," *RSC Adv.*, vol. 9, no. 14, pp. 7673–7679, 2019.
- [25] E. J. Bjerrum, M. Glahder, and T. Skov, "Data augmentation of spectral data for convolutional neural network (cnn) based deep chemometrics," *arXiv preprint arXiv:1710.01927*, 2017.
- [26] K. Seddiki, P. Saudeumont, F. Precioso, N. Ogrinc, M. Wisztorski, M. Salzet, I. Fournier, and A. Droit, "Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification," *Nat. Commun.*, vol. 11, no. 1, pp. 1–11, 2020.
- [27] Y. Luo and B.-L. Lu, "Eeg data augmentation for emotion recognition using a conditional wasserstein gan," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2018, pp. 2535–2538.
- [28] Z. Kong, J. Guo, A. Li, and C. Liu, "Physgan: Generating physical-world-resilient adversarial examples for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14254–14263.
- [29] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 13001–13008.
- [30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, Jun. 2019, pp. 4171–4186.
- [31] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Process Lett.*, vol. 51, no. 2, pp. 2007–2028, 2020.
- [32] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536–2544.
- [33] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 85–100.
- [34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [35] G. Kang, X. Dong, L. Zheng, and Y. Yang, "Patchshuffle regularization," *arXiv preprint arXiv:1707.07103*, 2017.
- [36] H. Inoue, "Data augmentation by pairing samples for images classification," *arXiv preprint arXiv:1801.02929*, 2018.
- [37] C. Summers and M. J. Dinneen, "Improved mixed-example data augmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*. IEEE, 2019, pp. 1262–1270.
- [38] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.
- [39] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3733–3748, 2021.
- [40] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
- [41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.
- [42] H. Gao, Z. Shou, A. Zareian, H. Zhang, and S.-F. Chang, "Low-shot learning via covariance-preserving adversarial augmentation networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 975–985.
- [43] E. Schwartz, L. Karlinsky, J. Shtok, S. Harary, M. Marder, A. Kumar, R. Feris, R. Giryes, and A. Bronstein, "Delta-encoder: an effective sample synthesis method for few-shot object recognition," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 2850–2860.
- [44] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7278–7286.
- [45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 3020–3029.
- [46] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

- [47] S. Jenni and P. Favaro, "Self-supervised feature learning by learning to spot artifacts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2733–2742.
- [48] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1–8.
- [49] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 29, 2016.
- [50] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [51] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2017, pp. 1126–1135.
- [52] X. Shen, *Spectroscopy and optical properties of semiconductors*. Beijing, China: China Science Publishing, 2002.
- [53] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [54] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.
- [55] J. Liu, D. Zhang, D. Yu, M. Ren, and J. Xu, "Machine learning powered ellipsometry," *Light Sci. Appl.*, vol. 10, no. 1, pp. 1–7, 2021.
- [56] J. Wei, G. Zhu, Z. Fan, J. Liu, Y. Rong, J. Mo, W. Li, and X. Chen, "Genetic u-net: automatically designed deep networks for retinal vessel segmentation using a genetic algorithm," *IEEE Trans. Med. Imaging*, vol. 41, no. 2, pp. 292–307, 2021.
- [57] M. Fukuhara, K. Fujiwara, Y. Maruyama, and H. Itoh, "Feature visualization of raman spectrum analysis with deep convolutional neural network," *Anal. Chim. Acta*, vol. 1087, pp. 11–19, 2019.
- [58] B. Lafuente, R. T. Downs, H. Yang, and N. Stone, "The power of databases: The ruff project," in *Highlights in mineralogical crystallography*. De Gruyter (O), 2015, pp. 1–30.
- [59] R. Kokaly, R. Clark, G. Swayze, K. Livo, T. Hoefen, N. Pearson, R. Wise, W. Benzel, H. Lowers, R. Driscoll *et al.*, "Usgs spectral library version 7 data: Usg geological survey data release," *United States Geological Survey (USGS): Reston, VA, USA*, 2017.
- [60] W. Shifu and X. Yizhuang, *Fourier transform infrared spectroscopy*. Beijing, China: Chemical Industry Press, 2016.
- [61] A. I. Apopei, N. Buzgar, and A. Buzatu, "Raman and infrared spectroscopy of kaersutite and certain common amphiboles," *Analele Stiintifice de Universitatii AI Cuza din Iasi. Sect. 2, Geologie*, vol. 57, no. 2, p. 35, 2011.
- [62] W. R. Ryall and I. M. Threadgold, "Evidence for [(SiO3)5] ∞ type chains in inosite as shown by x-ray and infrared absorption studies," *Am. Mineral.*, vol. 51, no. 5-6, pp. 754–761, 1966.
- [63] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.



Yingying Wu received the B.Sc. degree in computer science and technology from Nankai University, Tianjin, China, in 2021. She is currently pursuing the M.Sc. degree in artificial intelligence with the Institute of Robotics and Automatic Information System (IRAIS), College of Artificial Intelligence, Nankai University, Tianjin, China. Her current research interests include spectroscopic data analysis, machine learning and computer vision.



Jinchao Liu received the B.Sc. degree in automation and M.Sc. degree in control science and engineering from the Wuhan University of Technology, Wuhan, China, in 2004 and 2007, respectively, and the Ph.D. degree in mechanical engineering from the Technical University of Denmark, Kgs. Lyngby, Denmark, in 2012. He also holds an Industrial Ph.D. certificate conferred by the Danish Ministry of Science, Innovation and Higher Education. He is currently an Associate Professor with the Institute of Robotics and Automatic Information System (IRAIS), College of Artificial Intelligence, Nankai University, China. His current research interests include machine learning, computer vision, robotics and AI for science.



Yan Wang received the B.Sc. degree in computer science and technology from the Beijing Jiao Tong University, China in 2001. He obtained the M.Sc. degree in Information Technology and the Ph.D. degree in computer science from the University of Nottingham in 2003 and 2009 respectively. He then worked as a postdoctoral researcher at University of Nottingham and University of Bradford respectively between 2009 and 2013. He is a senior scientist and software engineer at VisionMetric Ltd., UK. Dr. Wang's research interests include machine learning, computer vision and medical imaging. He is specialized in facial image related applications include face modeling, recognition and facial image generation and editing.

search interests include machine learning, computer vision and medical imaging. He is specialized in facial image related applications include face modeling, recognition and facial image generation and editing.



Stuart Gibson is a Senior Lecturer in the School of Physics and Astronomy at the University of Kent, having been appointed as an academic staff member in 2007. He is the co-inventor of the EFIT-V facial identification system which is currently used by the majority of UK police constabularies and in numerous other countries. The main theme of Dr. Stuart Gibson's research is novel applications of machine learning and image processing in the natural sciences.



Margarita Osadchy received the Ph.D. degree (Hons.) in computer science from the University of Haifa, Haifa, Israel, in 2002. She was a Visiting Research Scientist with the NEC Research Institute, Princeton, NJ, USA, and then a Post-Doctoral Fellow with the Department of Computer Science, Technion, Haifa. She is currently an Associate Professor with the Department of Computer Science, University of Haifa. She is also a member of the Data Science Research Center and the Scientific Committee of the Center for Cyber Law and Policy, University of Haifa. Her main research interests are deep learning, machine learning, computer vision, and computer security and privacy.

terests are deep learning, machine learning, computer vision, and computer security and privacy.



Yongchun Fang (Senior Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in control theory and applications from Zhejiang University, Hangzhou, China, in 1996 and 1999, respectively, and the Ph.D. degree in electrical engineering from Clemson University, Clemson, SC, USA, in 2002. He is currently a Professor with the Institute of Robotics and Automatic Information System (IRAIS), College of Artificial Intelligence, Nankai University, Tianjin, China. His main research interests include nonlinear control, control of underactuated systems, visual servoing, intelligent systems and AFM-based nanosystems. Prof. Fang received the National Science Fund for Distinguished Young Scholars of China.

terests include nonlinear control, control of underactuated systems, visual servoing, intelligent systems and AFM-based nanosystems. Prof. Fang received the National Science Fund for Distinguished Young Scholars of China.