An Adversarial Approach to Discriminative Modality Distillation for Remote Sensing Image Classification

Shivam Pande[1]* shivam_pande@iitb.ac.in
Avinandan Banerjee[2]* avinandanbanerjee99@gmail.com
Saurabh Kumar[1]* saurabhkm@iitb.ac.in
Biplab Banerjee[1] getbiplab@gmail.com
Subhasis Chaudhuri[1] sc@iitb.ac.in
[1]Indian Institute of Technology Bombay, Mumbai, India
[2]Jadavpur University, Kolkata, India

Abstract

We deal with the problem of modality distillation for the purpose of remote sensing (RS) image classification by exploring the deep generative models. From the remote sensing perspective, this problem can also be considered in line with the missing bands problem frequently encountered due to sensor abnormality. It is expected that different modalities provide useful complementary information regarding a given task, thus leading to the training of a robust prediction model. Although training data may be collected from different sensor modalities, it is many a time possible that not all the information are readily available during the model inference phase. This paper tackles the problem by proposing a novel adversarial training driven hallucination architecture which is capable of learning discriminative feature representations corresponding to the missing modalities from the available ones during the test time. To this end, we follow a teacher-student model where the teacher is trained on the multimodal data (learning with privileged information) and the student model learns to subsequently distill the feature descriptors corresponding to the missing modality. Experimental results obtained on the benchmark hyperspectral (HSI) datasets and another dataset of multispectral (MS)-panchromatic (PAN) image pairs confirm the efficacy of the proposed approach. In particular, we find that the student model is consistently able to surpass the performance of the teacher model for HSI datasets.

1. Introduction

The current generation has witnessed the accumulation of a large volume of satellite remote sensing (RS) images, thanks to a number of successful space missions. These images play an extremely important role in applications concerning urban mapping, disaster management, city planning, to name a few [30]. Different types of RS images are capable of capturing completely diverse aspects regarding the underlying Earth's surface: from detailed spatial (VHR optical) to high spectral information (hyperspectral). Hence, it is needless to mention that the joint exploration of different RS modalities is worthwhile for improved Earth monitoring. However, it is not always possible to obtain all the cross-modal information together in time-critical situations (concerning the facts related to different temporal resolutions for the sensors, quick response for disaster management etc.). This further instigates the following scenario in terms of developing a machine learning system: *given that the training data are captured offline, it is always possible to train a model with multimodal information, however, the test samples may not always be available with all the modalities on the fly*.

Nonetheless, it is always preferable to train prediction models with extra (multi-modal) information: a paradigm known as *learning with privileged information* [22]. However, if the test samples are not consistent with the training data in terms of the feature dimensions, the trained model apparently cannot be evaluated on the test set. There are two possible solutions in this regard: i) use the information that are deemed to be available both during training and testing, however, compromising the performance of the learning model to some extent, or ii) train the model with the privileged side information and device a way to approximate the missing information of the test data given the available information, a phenomenon generally known as *modality distillation through hallucination* [9]. In this paper, we propose a solution to the distillation problem in the area of RS image classification by exploring deep gen-

*Authors contributed equally

erative model driven teacher-student architecture. While the teacher model is trained with all the modalities, we subsequently train a student model that work on the available and hallucinated fetures. We consider two experimental scenarios: i) RS scene classification using multispectral (modality-1) and panchromatic (modality-2) data, and ii) HSI classification where two non-overlapping subsets of the spectral bands denote both the concerned modalities. Note that discriminative feature descriptors are initially learned specific to both the modalities and the distillation module is devoted to learn the feature mapping from modality-1 to modality-2.

There has recently been a surge in this type of knowledge transfer in the area of computer vision [10, 13] within the teacher-student based distillation framework. On the other hand, the problem, though of immense importance in the field of RS, has hardly been studied to date. To the best of our knowledge, the only noteworthy endeavor [15] in this respect has directly extended the work of [13] to support RS images. However, the reconstruction loss based method of [13] does not always learn well the overall data distributions. Another important issue in this regard is the discriminativeness of the modality specific feature descriptors. A conditional generative adversarial network (C-GAN) based model has recently been utilized in this respect [31]. However, the vanilla GAN models with binary discriminator frequently get affected by the mode collapse problem during the learning phase of hallucination.

In this paper, we intend to tackle both the aforementioned issues of discriminative modality distillation problem under the realm of teacher-student based network. The teacher network is designed as a multi-stream network with a multi-layer classifier where each stream focuses on learning discriminative feature representations corresponding to the specific modality under consideration. This is followed by a C-GAN based hallucination model to generate the features of the absent modality conditioned on the features corresponding to the available modality. Here we propose to consider $2C$ number of nodes for C classes in the C-GAN discriminator and carry out the min-max type optimization in order to ensure discriminativeness of the hallucinated features. Finally, we design the student network that takes the features corresponding to the available and hallucinated modalities and subsequently carry out the C-class classification task. This setup, apart from mitigating the mode collapse problem, ensures discriminativeness in the hallucinated features. In addition, we utilize the soft-target based knowledge distillation (KD) paradigm for training the student's classifier. We summarize our major contributions as:

- We introduce a novel teacher-student based modality distillation framework for RS image classification where a novel C-GAN based cross-modality mapping module is proposed. We also consider the KD technique to ensure that the student's classifier does not diverge too much from the teacher's classifier.
- We perform data augmentation through noise perturbation on the teacher's training samples in order to train the hallucination and student models.
- We perform extensive experiments on HSI classification and RS scene classification using MS-PAN image pairs where improved results can be observed.

2. Related works

The two concepts that form the backbone of this research are *learning under privileged information* and *modality distillation*. We discuss about the related prior endeavors in the following.

2.1. RS image classification

RS data may contain images from several modalities such as multispectral, synthetic aperture radar (SAR), light detection and ranging (LiDAR), panchromatic images and hyperspectral imageries (HSI) [45]. HSI classification and analysis is one of the most sought topics in the field of remote sensing owing to the high dimensionality and information content of the HSIs. Several research works emanating from the field of conventional machine learning [23, 18, 38] or deep learning [27, 24, 25] have focused on fast and efficient classification of HSIs from both spectral and spatial perspectives. However, the research is not just limited to conventional classification paradigm but also branches into other areas such as domain adaptation [3, 29, 37] and zero shot learning [33, 17, 20].

In contrast to working with a single modality, there have been several approaches to combine different RS image modalities in synergistic way to get maximum results from the same feature set. This is obtained through image fusion and multi-modal learning. Recently, the deep learning techniques are used to combine the learned features corresponding to different modalities in a principled manner. [21] proposed an architecture called *Pan-Sharpening* GAN (PSGAN) to fuse panchromatic image and multispectral images of a given geographical area. Similarly, [12] proposed a method to fuse SAR imageries (from Sentinel-1) and multispectral imageries (from Sentinel-2) using a model composed of CNN with residual connections and C-GAN. [1] introduced a multi-modal segmentation model called *OrthoSeg* that works on three modalities, namely RGB images, infrared images and digital surface model (DSM). [4] came up with a modification over existing CNNs (based on *Squeeze and Excitation Networks* [14]) to fuse LiDAR data and HSI data. The model comprises of parallel streams for each modality while a residual block is used in each stream to extract hierarchical and multi-scale features.

2.2. Learning under privileged information (LUPI)

The LUPI paradigm was first introduced in [35] specifically for the support vector machines (SVM) classifiers. This idea was later extended in [28] where the authors introduce a concept of privileged empirical risk minimization to identify a faster learning function in decision space. Modifications were incorporated into existing LUPI paradigm in [34] while it was used along with knowledge distillation in [22]. LUPI has subsequently been implemented in several other domains such as unsupervised learning [5], metric learning [8, 39, 7], object localization [6], face detection [40], expression recognition [36] and many more. However, all these works incorporate privileged information in the conventional machine learning setting where the idea is either to maximize the margin among the classes or maximize the likelihood of an instance belonging to a certain class. More recently, [43] introduces the idea of LUPI in deep learning setting where instead of teacher-student framework, an ensemble of students are considered, which encourages co-operative learning. This is achieved by incorporating two losses namely a supervised learning loss and mimicry loss. The latter one tries to match the posterior of each student network to the class probabilities of other students.

2.3. LUPI with modality distillation

The idea of LUPI is further extended to deep learning setting mostly in combination with knowledge distillation frameworks. [2] incorporated LUPI with CNN for image categorization where the mapping difference between the visual feature and word embeddings is used as privileged information (termed as *privileged cost*). [26] introduced a multimodal method for gesture recognition from video dataset by combining multiscale learning (using spatial and temporal scales) coupled with multimodal learning. A multimodal CNN is employed to fuse the different modalities and perform classifcation. In addition, a regularization technique called ModDrop is presented, where during the fusion, weights corresponding to certain modalities are dropped for each iteration based on probabilities from Bernoulli's selector. [13] uses depth as extra information in a convolutional RGB object detection framework along with modality hallucination. [9] improvised the aforementioned method by incorporating adversarial learning in hallucination process where depth is used along with RGB images during training phase but omitted during testing phase. [31] proposed a C-GAN based approach to generate the missing modalities from the available ones. LUPI is further used in several other deep learning based computer vision applications such as brain tumour detection [42], action recognition using RNNs [32] and multi-instance multi-label (MIML) learning [41].

Note that there exists only very few works for modality distillation in RS which are primarily the direct extension of the existing techniques to the RS space [15].

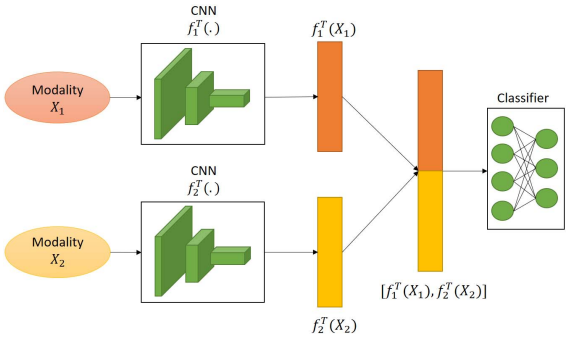

Figure 1. Schematic of teacher network. The two modalities X_1 and X_2 are fed to featurents $f_1^T(.)$ and $f_2^T(.)$. The obtained features are concatenated and sent to the classifier f_C^T.

2.4. How are we different?

The existing works closest to us are [31, 9] which also exploit the C-GAN architecture for modality distillation. However, i) while both the [31, 9] approaches are focused on the RGB-D based human activity recognition, ours is the first C-GAN based modality distillation framework in the domain of RS, ii) we have introduced a novel C-GAN discriminator architecture with $2C$ output nodes as opposed to the $C+1$ nodes mentioned in [31, 9] which can better deal with the mode collapse problem of GAN, while still preserving the discriminativeness of learned features. iii) While both the teacher and student classifiers of [9] are designed by averaging the softmax scores of the modality-specific classifiers, we, on the other hand, prefer learnable classifiers both in the teacher and student networks. We also add a KD term while training the student's classifier in a more constrained manner, and iv) we utilize the notion of data augmentation to encourage the training of the student's classifier on extra novel samples in comparison to the teacher model.

3. Proposed Methodology

We discuss the proposed algorithm in this section. The training pipeline is broadly divided into three phases: i) training of teacher network, ii) training of hallucination module and, iii) training of student's classifier practically to be utilized during inference. The stages are detailed in the following.

3.1. Training the teacher network

Let us consider a dataset $\mathcal{X} = \{x_1^i, x_2^i, y^i\}_{i=1}^N$ where $x_1^i \in \mathcal{X}_1$ and $x_2^i \in \mathcal{X}_2$ define the two modality specific inputs and $y^i \in \mathcal{Y}$ denotes the class labels from a pre-defined

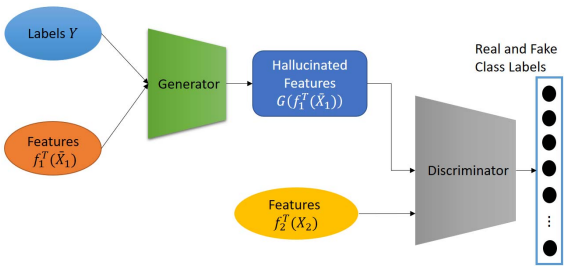

Figure 2. Schematic of feature hallucination through C-GAN. Extracted features $f_1^T(\bar{\mathcal{X}}_1)$ and labels are fed to the generator G that hallucinates features $G(f_1^T(\bar{\mathcal{X}}_1))$, which are fed to the discriminator D along with $f_2^T(\mathcal{X}_2)$. D is then trained against the real and fake sets of classes.

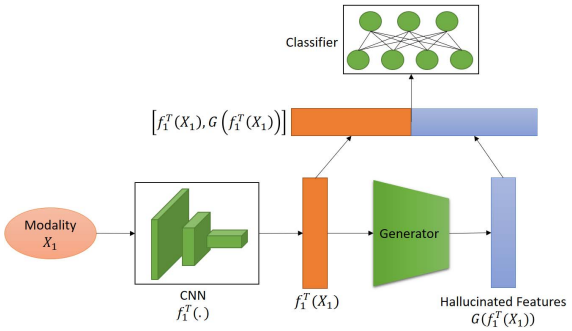

Figure 3. Schematic of student network. Modality X_1 is fed to $f_1^T(.)$. Extracted features $f_1^T(\mathcal{X}_1)$ are then sent to the trained G to hallucinate the missing modality through $G(f_1^T(\mathcal{X}_1))$. Both $f_1^T(\mathcal{X}_1)$ and $G(f_1^T(\mathcal{X}_1))$ are concatenated and sent to the classifier f_C^S.

set of C land-cover classes. For the MS-PAN data, we consider the images as the inputs for both the modalities while for the HSI datasets, local patches centered around the pixel locations are considered in order to jointly capture both the spectral and spatial information. The teacher model is regarded as the cumbersome model which is trained with the privileged information. In particular, the teacher network $\mathcal{T}$ consists of two modality specific feature extractors $f_1^T(\cdot)$ and $f_2^T(\cdot)$ (we refer to them as featurenets) and a classification model f_C^T which considers the concatenated feature representations ($\tilde{x}^i = [f_1^T(x_1^i) f_2^T(x_2^i)]$) obtained from the featurenets as inputs and maps them onto the label space. We realize both the feature extractors in terms of convolutional networks while the classification module is defined in terms of a feed-forward neural network model. Classification is subsequently carried out in terms of the softmax cross-entropy loss as follows:

$$\mathcal{L}_T = -\mathbb{E}_{(x_1^i, x_2^i, y^i)\in\mathcal{X}}[y^i \log f_C^T(\tilde{x}^i)] \quad (1)$$

The schematic for the teacher network can be found in Fig. 1.

3.2. Modality hallucination using C-GAN

In our experimental setting, modality $\mathcal{X}_2$ is assumed to be absent during inference and $\mathcal{X}_1$ as the available modality. Hence, our hallucination model $\mathcal{H}$ is devoted to hallucinate $f_2^T(x_2^i)$ given $f_1^T(x_1^i)$ as the input. Note that we are interested in hallucinating the features as learned by $\mathcal{T}$ than the actual data themselves considering the discriminativeness of the features learned by $f_2^T(\cdot)$. By design, the C-GAN model consists of a feature generator (G) or the hallucination stream and a discriminator (D) network. The feature generating network is conditioned on the samples from $f_1^T(\mathcal{X}_1)$ along with the label vector $\mathcal{Y}$, respectively. Moreover, we introduce two intuitive modifications in the C-GAN architecture so as to avoid any possible trivial solution and to ensure the generation of more discriminative hallucinated features which are at par with the features learned by $\mathcal{T}$. They are:

- We consider samples from $\mathcal{X}$ as well as augmented samples by adding random Gaussian noise $z \in \mathcal{N}(0,1)$ to $f_1^T(\mathcal{X}_1)$. We note that for a given (x_1^i, x_2^i, y^i), even if we perturb $f_1^T(x_1^i)$ to generate a cloud of feature points surrounding $f_1^T(x_1^i)$, we do not modify $f_2^T(x_2^i)$ while hallucinating in order to ensure robustness of the cross-modality mapping.

- The discriminator D is designed to output $2C$ class scores where the *real* and *fake* samples are considered per class basis. This mitigates the generation of potentially spurious samples for modality $f_2^T(\mathcal{X}_2)$. In particular, the label outputs of D (denoted as $\bar{\mathcal{Y}}$) are encoded as vectors of length $2C$ where the given j^{th} and $j+1^{th}$ index are labeled as 1 each based on whether the input sample is the feature descriptor corresponding to a real sample from the $j^{th} \in \mathcal{Y}$ category from $\mathcal{X}_2$ or hallucinated (fake).

Let $\bar{\mathcal{X}} = (\bar{x}_1^j, x_2^j, \bar{y}^j)_{j=1}^M$ where $\bar{x}_1^j$ is either a sample obtained from $\mathcal{X}$ (referred to as x_1^j) or a perturbed version of x_1^j ($x_1^j + z$) and $\bar{y}^j \in \bar{\mathcal{Y}}$, be the training samples used to train $\mathcal{H}$. To this end, both the G and D are trained based on the adversarial min-max strategy through the optimization of $\mathcal{L}_{hal}$ as stated in the equation 2.

$$\min_G \max_D \mathcal{L}_{hal} = \mathbb{E}_{(\bar{x}_1^j, \bar{y}^j, y^j)}[\log D(G(f_1^T(\bar{x}_1^j | y^j, \bar{y}^j)))] + \mathbb{E}_{(x_2^j, \bar{y}^j, y^j)}[\log D(f_2^T(x_2^j), \bar{y}^j)] \quad (2)$$

where $|$ denotes the concatenation operation. Fig. 2 shows the schematic of modality hallucination through C-GAN.

3.3. Training the student network

For the student model $\mathcal{S}$, the featurenet $f_1^T(\cdot)$ for the available modality $\mathcal{X}_1$ from the teacher $\mathcal{T}$ and the trained generator of the hallucination network $\mathcal{H}$ are kept fixed while only the student's classifier is trained. Since modality $\mathcal{X}_2$ is absent during the testing phase, the student network $\mathcal{S}$ is trained using the availability modality $\mathcal{X}_1$ only. In principle, $\bar{x}_1^j$ is first sent to featurenet $f_1^T(.)$ to obtain the feature representation $f_1^T(\bar{x}_1^j)$ which is subsequently forwarded to the generator G in order to generate the hallucinated features $G(f_1^T(\bar{x}_1^j))$ corresponding to the missing modality. Next, $f_1^T(\bar{x}_1^j)$ and $G(f_1^T(\bar{x}_1^j))$ are concatenated as $\hat{x}^j = [f_1^T(\bar{x}_1^j), G(f_1^T(\bar{x}_1^j))]$ and is further used to train the student's classifier f_C^S. However, we require that the predictions of f_C^S should not diverge much from the predictions on f_C^T on similar samples. In order to ensure the same, we consider to jointly optimize a cross-entropy based classification loss and a knowledge distillation loss between the teacher and the student which can be mentioned as:

$$\mathcal{L}_S = -\mathbb{E}_{(\hat{x}^j, y^j)}[y^j \log f_C^S(\hat{x}^j)] + \lambda(||\mathbf{q}_T^j - \mathbf{q}_S^j||_2) \quad (3)$$

where, $\mathbf{q}_T^j$ and $\mathbf{q}_S^j$ are the softmax probability vectors of size $C \times 1$ (where C is the number of classes) for j^{th} sample from teacher and student networks, respectively, where a high temparature value is considered within the softmax formulation. Precisely, the temperature based softmax normalization is given as:

$$q_c^j = \frac{e^{z_c^j/T}}{\sum_{c=1}^{C} e^{z_c^j/T}} \quad (4)$$

where, q_c^j is the more soften softmax probability for c^{th} class, z_c^j is the logit scores for c^{th} class given $\hat{x}^j$.

The hyperparameter λ is set empirically as we seek to give more weightage to the classification term given the teacher also performs some misclassification. The temperature is fixed to a higher value (> 1) while training the teacher and the student networks, however during testing of the student network, its value is set back to 1 (Fig. 3).

3.4. Inference

For a given test sample x in modality-1, we utilize the featurenet f_1^T and the hallucination generator G in order to obtain $f_1^T(x)$ and $G(f_1^T(x))$, respectively. These features are concatenated and fed to f_C^S in order to obtain the class label for the sample.

4. Experiments and Results

Extensive experiments are conducted on remote sensing datasets to validate the performance of our model and compare it to existing approaches. We detail the evaluations and discussions in the following.

4.1. Datasets

We consider two benchmark HSI datasets and another multimodal dataset for RS where a large number of multispectral-PAN patches are provided in pairs. While we solve the pixel classification problem for the HSI datasets by exploring both the spatial and spectral information, we deal with the problem of scene recognition for the MS-PAN dataset.

Multispectral-panchromatic dataset: The dataset consists of multispectral (MS) imagery of 4 bands and panchromatic (PAN) imagery for a given geographical area with a total of 80000 image pairs emanating from eight land-cover classes. All the imageries are collected from multispectral and panchromatic sensors of GF-1 satellite [19]. The size of each MS image is $64 \times 64 \times 4$ with 2m spatial resolution whereas the size of the corresponding PAN imagery is 256×256 pixels with spatial resolution of 8m. Fig. 4 shows colour composite for one of the MS samples and its panchromatic counterpart.

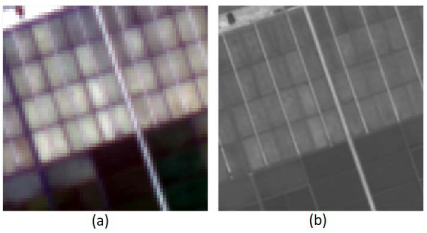

Figure 4. Houston hyperspectral dataset: (a) Colour composite of three bands for MS image. (b) PAN image.

Indian Pines hyperspectral dataset: Acquired by AVIRIS sensor over the North-western Indiana, this HSI dataset [44] is captured in 200 bands each of size 145×145 with spatial resolution of 20m. The area contains pixels from sixteen pre-defined set of land-cover classes. As a whole, there exist a total of 10249 pixels in the scene with associated ground-truth labels (fig. 5 [44]).

Houston hyperpectral dataset: The imagery for Houston dataset [11] (Fig. 6) is collected by National Center for Airborne Laser Mapping over the campus of University of Houston and surrounding urban areas. The imagery comprises of 144 bands each of size 1905×349 and a total of 15 land-use/land-cover classes with spatial resolution of 2.5m. The number of pixel vectors with associated ground truth classes is 15029.

4.2. Model Architectures

In this section, we detail the experimental protocols followed and the model architectures for $\mathcal{T}$, $\mathcal{H}$, and $\mathcal{S}$, respec-

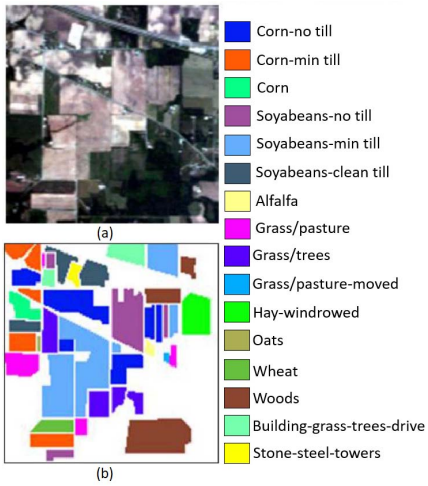

Figure 5. Indian pines hyperspectral dataset: (a) Colour composite of three bands from red, green and blue wavelengths. (b) Groundtruth image.

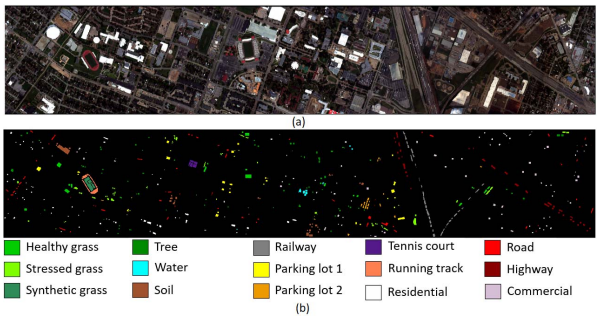

Figure 6. Houston hyperspectral dataset: (a) Colour composite of three bands from red, green and blue wavelengths. (b) Groundtruth with classes.

tively for each of the datasets.

Teacher model: The teacher network consists of two parallel featurenets which are realized in terms of CNN architectures consisting of convolutional, pooling, and fully-connected layers. In MS-PAN dataset, the MS images are fed to featurenet-1 while the PAN images are fed to featurenet-2, respectively. Each of the MS images is of size $64 \times 64 \times 4$ while each PAN image is 256×256, therefore, the CNN corresponding to MS image is fixed to 3 layers of convolutions and pooling while that corresponding to PAN image comprises of 4 convolutional and pooling layers since the latter requires further reduction in size. For both the CNN encoders, 128 filters of size 5×5 are used in their first layers, the second layer uses 128 filters of size $5 \times 5 \times 128$ while the third layer uses 64 filters each of size $5 \times 5 \times 128$. The fourth layer of featurenet-2 comprises of 64 filters with size $5 \times 5 \times 64$. Both the layers use *same padding* and *max-pooling* followed by *batch normalization* and *dropout* with a rate of 10% and *ReLU* activation function. The strides in both convolutional filters and pooling kernels is kept as 1.

In hyperspectral datasets, each pixel is a $1 \times 1 \times B$ pixel vector (where B is the number of bands). Hence, a patch of size 17×17 is created centering around each pixel location. Thereafter, patches from each modality are sent to corresponding CNN encoder for training. The CNN encoders are similar to those used for MS-PAN dataset upto the third layer.

The output from each CNN feature extractor is flattened and sent to a two layer fully connected (FC) encoder with 512 and 200 nodes in the first and second layers, respectively, to reduce the dimension of the features to 200. The outputs from the FC encoder are concatenated and sent to a 3 layer FC neural network classifier $f_C^T(.)$ that consists of 400 and 200 nodes in the first and second layers followed by a softmax layer with C nodes where C is the number of classes. The classifier is trained on categorical cross entropy loss.

Hallucination model: In the C-GAN based hallucination model, both generator and discriminator are fully connected 3 layer neural networks with a softmax layer at the end of the latter. However, the softmax layer consists of $2C$ nodes since it takes into account both real and fake samples per class. Both the networks use *batch normalization*, *dropout* with 10% rate and *ReLU* activation.

Student model: The student model consists of the featurenet corresponding to the available modality, the trained generator (weights of both are kept fixed) and an FC multilayer classifier. The classifier is a 3 layer fully connected neural network $f_C^S(.)$ that mimics the architecture of $f_C^T(.)$ and is trained on the extracted features from available modality and generated features for absent modality.

All the models are trained using *Adam* optimizer [16] with a learning rate of 0.001.

4.3. Discussions

We consider two baselines for evaluating the performance of the student's classification module: i) we train separate classification networks for each of the modalities and report the classification performances, and ii) the performance of the teacher model where the modality specific features are fused to train the teacher's classifier. For baseline comparison, we consider a randomly selected set of 25% samples to train the student-teacher model and the remaining 75% samples are utilized during testing. Specific to HSI, we consider two sets of randomly selected non-overlapping bands for constituting modality-1 and 2, respectively.

Apart from the baselines, we compare our model to the hallucination techniques inspired from [9] (where the hallucination is carried out using an encoder) and [10] (where

Network	Accuracy (in %)
Stream MS	90.17
Stream PAN	92.58
Teacher	95.46
Teacher (avg)	98.07
Student with MS absent [9]	81.06
Student with PAN absent [9]	81.06
Student with MS absent [10]	37.24
Student with PAN absent [10]	37.35
Student with MS absent (proposed)	82.75
Student with PAN absent (proposed)	86.40

Table 1. Results on MS-PAN dataset.

Network	Accuracy (in %)
Stream 1	73.53
Stream 2	54.91
Teacher	70.28
Student (proposed)	80.57

Table 2. Results on Indian Pines dataset.

Network	Accuracy (in %)
Stream 1	95.28
Stream 2	91.80
Teacher	98.17
Student (proposed)	97.96

Table 3. Results on Houston dataset.

the softmax probabilities from two streams are averaged before classification) since the two aforementioned techniques have been state of the art in the areas of modality hallucination. The results for baseline comparison for MS-PAN dataset, Indian pines dataset and Houston dataset are tabulated in tables 1, 2 and 3 respectively.

We also carry out sensitivity analysis on Indian pines dataset by varying the ratio of bands in each modality and temperature. In addition, our C-GAN with discriminator trained on $2C$ classes is also compared against C-GANs with binary discriminator and with discriminator having $C+1$ classes.

MS-PAN dataset: For MS-PAN dataset, it is observed that the accuracies for the MS stream (90.17%) and PAN stream (92.58%) are comparable and that of the $\mathcal{T}$ (95.46%) surpasses the above two, which was expected. For $\mathcal{S}$, the accuracies recorded for hallucination of PAN imagery (86.40%) as well as for MS imagery (82.75%) are less than both streams as well as teacher network. The similar trend is observed for [9] as well where hallucination of PAN imagery from MS and vice versa give an accuracy of 81.06% each. It could be concluded from these observations that the MS and PAN bands do not share much correlated information and hence features from one modality are not efficiently

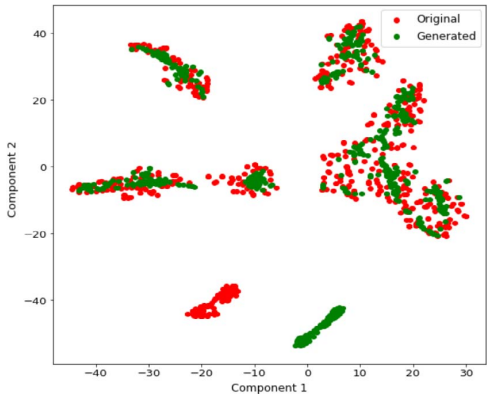

Figure 7. t-SNE comparing the original and encoder generated features for PAN imagery.

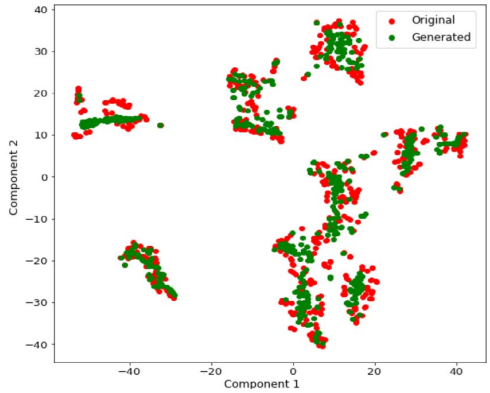

Figure 8. t-SNE comparing the original and encoder generated features for MS imagery.

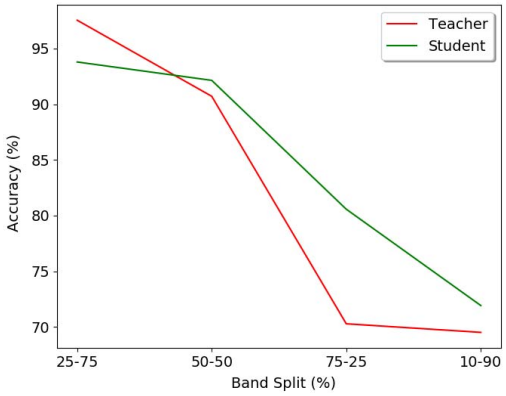

Figure 9. Variation in test accuracy with change in modality split ratios conducted on Indian pines dataset.

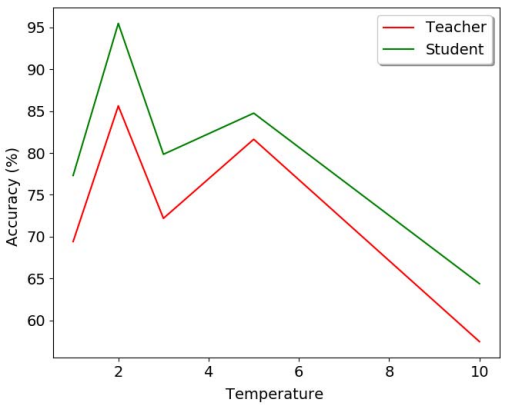

Figure 10. Variation in test accuracy with change in temperature conducted on Indian pines dataset.

generating the features of other modality. However, while using the technique based on [10], it is observed that even though the $\mathcal{T}$ gives accuracy as high as 98.07%, the student networks give accuracies as low as 37.24% (when MS is hallucinated) and 37.35% (when PAN is hallucinated). This could mean that the CNN weights obtained as a result of averaging do not extract as informative features that could be used to generate the features of absent modality. Fig. 7 and fig. 8 show the t-SNE plots to compare generated PAN features from available MS features and vice versa through hallucination framework based on [9]. It is visible from the t-SNEs that the generated features of absent modality and original features of the same effectively overlap each other, thus showing efficient hallucination.

Indian pines dataset: For Indian pines dataset, the accuracy of $\mathcal{T}$ (70.28%) is less than stream 1 (73.53%) by 3.25%. In addition, the accuracy for stream 2 (54.91%) is reported much less than stream 1. From this, it can be concluded that the bands in modality 2 carry irrelevant information, which hinders the performance of $\mathcal{T}$ as well. The accuracy for $\mathcal{S}$ (80.57%) surpasses both single stream as well as teacher models. This shows that the generated features from the $\mathcal{H}$ were able to fully capture the distribution of the original features and overcome the effects of band correlation thereby leading to better results.

Fig. 9 presents the trend of sensitivity analysis for the percentage of bands included and first and second modalities respectively for Indian pines dataset. For every instance except $75\% - 25\%$ split, the $\mathcal{S}$ outperforms the $\mathcal{T}$. From this, it is concluded that even with less number of available bands, $\mathcal{H}$ is able to generate better features.

Fig. 10 shows the variation in accuracies of $\mathcal{T}$ and $\mathcal{S}$ with respect to change in temperature keeping the train-test split fixed to $25\% - 75\%$ and band split to $50\% - 50\%$. The maximum accuracy is achieved when the temperature is fixed to 2.

Houston dataset: In case of Houston dataset, we can see that the $\mathcal{T}$ with 98.17% accuracy outperforms both stream 1 (95.28%) and stream 2 (91.80%) which is in concurrence with our expectations. The accuracy of $\mathcal{S}$ (97.96%) also surpasses the ones obtained from both the streams and only lags behind the $\mathcal{T}$ with a difference of 0.21%. It could be inferred from this, that the $\mathcal{H}$ is able to generate absent modality with greater accuracy and therefore the student classifier is able to learn efficient representations.

We do not compare the performance on HSI datasets to [9] and [10] because there, the student network either beats the teacher network or lags behind by a very small margin, which shows the efficiency of hallucination of absent modality.

Choice of discriminator architecture: We choose a discriminator that is trained on two sets of classes (real and fake) based on the intuition that training on two sets would make the model more susceptible to identify interclass and intraclass variances and reduce the misclassification rate among the generated features. This is an improvement over using the binary discriminator since it is unable to capture the intraclass variance. In addition, our discriminator also works better than the one that trains on $C + 1$ classes since the chances of misclassification among generated features is high in the latter's case. The models that had binary and $C + 1$ class discriminators gave classification accuracies as low as 30% that further strengthens our intuition.

5. Conclusions

We deal with the problem of modality distillation in the context of RS image classification. Given our data are represented by multiple modalities, we consider the scenario when all the modalities are available during training but some of the modalities are missing for the test data. We follow the standard teacher-student framework where the teacher model is trained on the multi-modal information. The student model, on the other hand has two tasks: i) to learn to hallucinate the missing modalities from the available one, and ii) train the student's classifier by using the available and hallucinated modalities. For modality hallucination, we propose a novel C-GAN based model which ensures the generation of classwise distinctive samples. We perform extensive experiments of different kinds on RS datasets to showcase the efficacy of our model. As opposed to the scenario where the student's classifier is trained on the training samples deployed to train the target, we are currently interested in exploring the notion of zero-shot knowledge distillation in this context where pseudo samples are learned to train the student.

Acknowledgement: The project is partially supported by SERB-ECRA (grant no. ECR/2017/000365).

References

[1] P. Bodani, K. Shreshtha, and S. Sharma. Orthoseg: A deep multimodal convolutonal neural network architecture for semantic segmentation of orthoimagery. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2018. 2

[2] K. Chen and Z. Zhang. Learning to classify fine-grained categories with privileged visual-semantic misalignment. *IEEE Transactions on Big Data*, 3(1):37–43, 2016. 3

[3] C. Deng, X. Liu, C. Li, and D. Tao. Active multi-kernel domain adaptation for hyperspectral image classification. *Pattern Recognition*, 77:306–315, 2018. 2

[4] Q. Feng, D. Zhu, J. Yang, and B. Li. Multisource hyperspectral and lidar data fusion for urban land-use mapping based on a modified two-branch convolutional neural network. *ISPRS International Journal of Geo-Information*, 8(1):28, 2019. 2

[5] J. Feyereisl and U. Aickelin. Privileged information for data clustering. *Information Sciences*, 194:4–23, 2012. 3

[6] J. Feyereisl, S. Kwak, J. Son, and B. Han. Object localization based on structural svm using privileged information. In *Advances in Neural Information Processing Systems*, pages 208–216, 2014. 3

[7] S. Fouad and P. Tiňo. Ordinal-based metric learning for learning using privileged information. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2013. 3

[8] S. Fouad, P. Tino, S. Raychaudhury, and P. Schneider. Incorporating privileged information through metric learning. *IEEE transactions on neural networks and learning systems*, 24(7):1086–1098, 2013. 3

[9] N. C. Garcia, P. Morerio, and V. Murino. Learning with privileged information via adversarial discriminative modality distillation. *arXiv preprint arXiv:1810.08437*, 2018. 1, 3, 6, 7, 8

[10] N. C. Garcia, P. Morerio, and V. Murino. Modality distillation with multiple stream networks for action recognition. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 6, 7, 8

[11] P. Ghamisi, E. Maggiori, S. Li, R. Souza, Y. Tarabalka, G. Moser, A. De Giorgi, L. Fang, Y. Chen, M. Chi, S. B. Serpico, and J. A. Benediktsson. Frontiers in Spectral-Spatial Classification of Hyperspectral Images. *IEEE geoscience and remote sensing magazine*, 6(3):10–43, Sept. 2018. This is a preprint, to read the final version please go to IEEE Geoscience and Remote Sensing Magazine on IEEE XPlore. 5

[12] W. He and N. Yokoya. Multi-temporal sentinel-1 and-2 data fusion for optical image simulation. *ISPRS International Journal of Geo-Information*, 7(10):389, 2018. 2

[13] J. Hoffman, S. Gupta, and T. Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016. 2, 3

[14] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[15] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen. Urban land cover classification with missing data modalities using deep convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(6):1758–1768, 2018. 2, 3

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] A. Li, Z. Lu, L. Wang, T. Xiang, and J.-R. Wen. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):4157–4167, 2017. 2

[18] J. Li, J. M. Bioucas-Dias, and A. Plaza. Hyperspectral image segmentation using a new bayesian approach with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10):3947–3960, 2011. 2

[19] Y. Li, Y. Zhang, X. Huang, and J. Ma. Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11):6521–6536, 2018. 5

[20] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, and R. Wang. Deep few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2290–2304, 2018. 2

[21] X. Liu, Y. Wang, and Q. Liu. Psgan: a generative adversarial network for remote sensing image pan-sharpening. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 873–877. IEEE, 2018. 2

[22] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015. 1, 3

[23] G. Mercier and M. Lennon. Support vector machines for hyperspectral image classification with spectral-based kernels. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*, volume 1, pages 288–290. IEEE, 2003. 2

[24] L. Mou, P. Ghamisi, and X. X. Zhu. Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3639–3655, 2017. 2

[25] A. Mughees and L. Tao. Efficient deep auto-encoder learning for the classification of hyperspectral images. In *2016 International Conference on Virtual Reality and Visualization (ICVRV)*, pages 44–51. IEEE, 2016. 2

[26] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015. 3

[27] M. Paoletti, J. Haut, J. Plaza, and A. Plaza. A new deep convolutional neural network for fast hyperspectral image classification. *ISPRS journal of photogrammetry and remote sensing*, 145:120–147, 2018. 2

[28] D. Pechyony and V. Vapnik. On the theory of learnining with privileged information. In *Advances in neural information processing systems*, pages 1894–1902, 2010. 3

[29] Y. Qin, L. Bruzzone, B. Li, and Y. Ye. Tensor alignment based domain adaptation for hyperspectral image classification. *arXiv preprint arXiv:1808.09769*, 2018. 2

[30] J. A. Richards and J. Richards. *Remote sensing digital image analysis*, volume 3. Springer, 1999. 1

[31] S. Roheda, B. S. Riggan, H. Krim, and L. Dai. Cross-modality distillation: A case for conditional generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2926–2930. IEEE, 2018. 2, 3

[32] Z. Shi and T.-K. Kim. Learning and refining of privileged information-based rnns for action recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3461–3470, 2017. 3

[33] G. Sumbul, R. G. Cinbis, and S. Aksoy. Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):770–779, 2017. 2

[34] V. Vapnik and R. Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of machine learning research*, 16(2023-2049):2, 2015. 3

[35] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009. 3

[36] S. Wang, B. Pan, H. Chen, and Q. Ji. Thermal augmented expression recognition. *IEEE transactions on cybernetics*, 48(7):2203–2214, 2018. 3

[37] Z. Wang, B. Du, Q. Shi, and W. Tu. Domain adaptation with discriminative distribution and manifold embedding for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 2019. 2

[38] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki. Random forest ensembles and extended multiextinction profiles for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1):202–216, 2018. 2

[39] X. Xu, W. Li, and D. Xu. Distance metric learning using privileged information for face verification and person re-identification. *IEEE transactions on neural networks and learning systems*, 26(12):3150–3162, 2015. 3

[40] H. Yang and I. Patras. Privileged information-based conditional regression forest for facial feature detection. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013. 3

[41] H. Yang, J. Tianyi Zhou, J. Cai, and Y. Soon Ong. Miml-fcn+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2017. 3

[42] F. Ye, J. Pu, J. Wang, Y. Li, and H. Zha. Glioma grading based on 3d multimodal convolutional neural network and privileged learning. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 759–763. IEEE, 2017. 3

[43] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 3

[44] Z. Zhang, E. Pasolli, M. M. Crawford, and J. C. Tilton. An active learning framework for hyperspectral image classification using hierarchical segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(2):640–654, 2015. 5

[45] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. 2