FusAtNet: Dual Attention based SpectroSpatial Multimodal Fusion Network for Hyperspectral and LiDAR Classification

Satyam Mohla[1†*] Shivam Pande[2*] Biplab Banerjee[2] Subhasis Chaudhuri[1]
[1]Deptt. of Electrical Engineering [2]Centre of Studies in Resources Engineering
IIT Bombay, Mumbai, India
{satyammohla, pshivamiitb, getbiplab}@gmail.com, sc@ee.iitb.ac.in

Abstract

With recent advances in sensing, multimodal data is becoming easily available for various applications, especially in remote sensing (RS), where many data types like multispectral imagery (MSI), hyperspectral imagery (HSI), LiDAR etc. are available. Effective fusion of these multisource datasets is becoming important, for these multimodality features have been shown to generate highly accurate land-cover maps. However, fusion in the context of RS is non-trivial considering the redundancy involved in the data and the large domain differences among multiple modalities. In addition, the feature extraction modules for different modalities hardly interact among themselves, which further limits their semantic relatedness. As a remedy, we propose a feature fusion and extraction framework, namely FusAtNet, for collective land-cover classification of HSIs and LiDAR data in this paper. The proposed framework effectively utilizses HSI modality to generate an attention map using "self-attention" mechanism that highlights its own spectral features. Similarly, a "cross-attention" approach is simultaneously used to harness the LiDAR derived attention map that accentuates the spatial features of HSI. These attentive spectral and spatial representations are then explored further along with the original data to obtain modality-specific feature embeddings. The modality oriented joint spectro-spatial information thus obtained, is subsequently utilized to carry out the land-cover classification task. Experimental evaluations on three HSI-LiDAR datasets show that the proposed method achieves the state-of-the-art classification performance, including on the largest HSI-LiDAR dataset available, University of Houston (Data Fusion Contest - 2013), opening new avenues in multimodal feature fusion for classification.

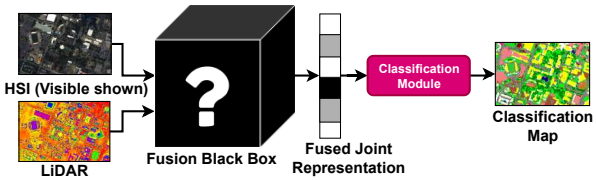

Figure 1. Generic schematic of a multimodal fusion based classification task. The objective is to effectively combine the two modalities (hereby HSI and LiDAR) such that the resultant representation has rich, fused features that are relevant and robust enough for accurate classification.

1. Introduction

With the advent of advanced sensing technologies, simultaneous acquisition of multimodal data for the same underlying phenomenon is possible nowadays. This is especially important in remote sensing (RS), owing to presence of satellite image data from several sources like multispectral (MSI), hyperspectral (HSI), synthetic aperture radar (SAR), panchromatic (PCI) sensors etc. as well as light detection and ranging (LiDAR), to name a few. Each source provides different kind of information about the same geographical region, which can aid in tasks relating to total scene understanding. For example, the detailed spectral information from HSI is commonly used to discriminate various materials based on their reflectance values, finding applications in agricultural monitoring, environment-pollution monitoring, urban-growth analysis, land-use pattern [1, 2]. Similarly, LiDAR data is used to obtain the elevation information, which is useful to distinguish objects within the same material [3]. Since the attributes of these modalities complement each other, they are extensively used in a cumulative fashion for multimodal learning in remote sensing domain [4, 5].

In the recent past, many ad-hoc and conventional techniques have been introduced for the fusion of HSI and LiDAR modalities due to their ability of digging the latent representations and features from the raw data [6, 7]. Besides, the concerned fusion strategies have been applied for

*Equal Contribution †Corresponding Author

different application scenarios as visible in [8, 9, 10, 11, 12], where conventional methods such as support vector machines (SVM), random forests (RF), rotation forests (RoF) etc. have been actively used for classification. [13] proposes an in flight fusion of LiDAR and HSI data. The intensity of HSI data is corrected with the help of cross-calibrated return intensity information obtained from airborne laser scanner (ALS).

Similarly, in the present era, deep learning is being actively used in the domain of multimodal fusion [14, 15]. The deep learning approach generally follows a multi-stream architecture where each stream corresponds to a single modality. These extracted features are then concatenated to be used as joint representation for further classification. Convolutional neural networks (CNNs) especially have been widely utilised as feature extractors in remote sensing community and shown to be more powerful than the conventional techniques for supervised inference tasks [16]. Although the multistream deep architectures have produced excellent performance measures, a key disadvantage to such an approach, however, is that the feature extraction of different modalities is carried out individually instead of utilising features from both modalities jointly. This causes some important shared high-level features from both the modalities to be missed out. Another key point is the fact that such a method may make different features significantly unbalanced, and the information may not be equally represented [17]. Given the multi-source feature embeddings, feature aggregation is an important stage. Simple concatenation or pooling of individual extracted features may have redundant information and thus the system might be prone to overfitting. Lastly, having large number of features by just concatenation may increase the dimensionality and due to lack of large labelled data, the model may suffer from curse of dimensionality [16]. Limited training samples or imbalanced data, along with the need of to avoid any human intervention in selecting features, have encouraged researchers to search for better joint feature learning methods.

Recently, the usage of attention learning mechanism has shown remarkable performance gain for different visual inference tasks [18, 19, 20]. Ideally, the attention modules highlight the prominent features while suppressing the irrelevant features through a self-supervised learning paradigm. However, in most of these research works, attention based learning is carried out only on a single modality and hence only similar kind of features are highlighted. Therefore, we are left with the task of designing such a network that takes the attention mask from one modality and use it to enhance the representations of other modality (Fig. 1). Based on this premise, the idea of multimodal attention is envisioned, where a complementing modality not only synergistically adds relevant information to the existing modality but also highlights such features that went "unnoticed" by the attention map derived from the existing modality. Inspired by these discussions, we propose FusAtNet, an attention based multimodal fusion network for land-cover classification given an HSI-LiDAR pair as input, as illustrated in Fig. 3. Our method involves extracting spectral features using "self-attention" in HSI and incorporate multimodal attention using the proposed "cross-attention" mechanism that uses LiDAR modality to derive an attention mask that highlights the spatial features of the HSI (Fig. 2). This interaction between spectral and spatial features leads to an intermediate representation which is further refined through self-attention based learning. This rich final representation is henceforth directed for classification. The key contributions are summarised as follows:

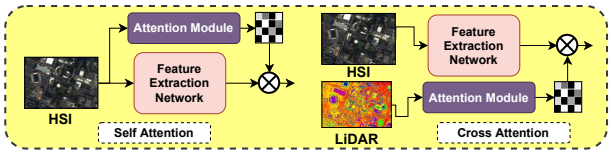

Figure 2. Self-attention vs cross-attention for multimodal fusion. The self-attention module (left) works only on single modality where both the hidden representations as well as the attention mask are derived from the same modality (HSIs). On the other hand, in the cross-attention module (right), the attention mask is derived from a different modality (LiDAR) and is harnessed to enhance the latent features from the first modality.

- To the best of our knowledge, ours is one of the first approaches to introduce the notion of attention learning for HSI-LiDAR fusion in the context of land-cover classification.
- In this regard, we introduce the concept of "cross-attention" based feature learning among the modalities, a novel and intuitive fusion method which utilises attention from one modality (here LiDAR) to highlight features in the other modality (HSI).
- We demonstrate state-of-the art classification performance on three benchmark HSI-LiDAR datasets outperforming all existing deep fusion strategies along with thorough robustness analysis.

2. Related work

By definition, the task of image fusion aims at synergistically combining images from different related modalities to generate a merged representation of the information present in the images, improving visual inference performance over the individual images. Growing interest from the multimedia community is reflected in various works like [21] where audio-visual crossmodal representation learning was proposed, in [22] where RGB-depth multimodal features were fused for scene classification and in shared cross modal image retrieval [23]. It is also an emerging topic in medical

image classification. For example, whereas [24] fuses information from MRI/PET, [25] utilises 4 different modalities utilising CNN as feature extractors for image segmentation. Unsupervised methods like [26] generate joint latent representation from data of different modalities using deep belief networks.

Remote sensing has been utilising several classical multimodal fusion methods such as decision fusion [27], kernel based fusion [28], PCA [29], intensity-hue-saturation (IHS) [30], wavelet based fusion methods [31] etc. for various applications in order to improve the classification performance of even the most conventional models.

In addition to classical methods, deep learning is being actively used in remote sensing field both for feature attention and multimodal learning. In feature attention, [19] presents a novel spectral-attention framework to highlight the reflectance characteristics of the hyperspectral image for better classification performance. Similarly, [20] introduces a spectral-spatial attention network using residual learning (to tackle vanishing gradient [32]) and convolution-deconvolution framework (to extract distinct spatial features) which collectively assist in robust classification. From the perspective of multimodal fusion in remote sensing, deep learning normally involves concatenation of extracted features from unimodal networks, and then sending them for classification. The entire model is trained in an end to end fashion. For example, [33] concatenated the Kronecker product of LiDAR derived features with the spectral features obtained from the HSI and used them for classification using a CNN model. [34] and [35] use a two stream model of image fusion where in one stream, a 3D-CNN is used to extract the spectral-spatial features from the HSIs while a 2D-CNN is used to extract the depth features from LiDAR dataset. It is important to note that the LiDAR data has been rasterised in the image domain as a digital elevation model (DEM) and digital surface model (DSM). The features are concatenated and sent to a deep neural network for fusion and finally classification. [36] proposed an adaptive technique of HSI-LiDAR fusion. Initially, the LiDAR and HSI features are extracted using a two-stream CNN where each stream corresponds to each modality. The streams follow the similar architecture and contains cascaded residual blocks (inspired from Face Alignment Network (FAN) [37] and hourglass networks [38]) to keep both the original and extracted features from fusion. The extracted features are then combined with original features using an adaptive technique based on squeeze and excitation networks [39] where, instead of simply concatenating the features, each feature is assigned a specific weight. The weighted tensors are flattened and concatenated and sent to a fully connected layer for classification.

As already mentioned, the existing techniques for HSI-LiDAR fusion overlook the aspect of attention based feature learning. On the other hand, FusAtNet incorporated different attention learning modules within its framework for better cross-modal feature extraction. Additionally, we introduce the notion of cross-modal attention which is a novel paradigm in the realm of feature fusion.

3. Proposed method

The objective of this work is to perform pixel based classification by harnessing the spectral and spatial information constituted in HSIs and the depth and intensity information encoded in LiDAR.

To accomplish this task, we consider HSI and LiDAR patches $\mathcal{X} = \{\mathbf{x}_H^i, \mathbf{x}_L^i\}_{i=1}^n$ that are centered around the ground truth pixels $\mathcal{Y} = \{y^i\}_{i=1}^n$. Here, $\mathbf{x}_H^i \in \mathbb{R}^{M\times N\times B_1}$ and $\mathbf{x}_L^i \in \mathbb{R}^{M\times N\times B_2}$ where, B_1 and B_2 denote the number of channels in HSI and LiDAR modalities respectively, while n denote the number of available groundtruth samples. The groundtruth labels $y_i^n \in \{1, 2, ..., K\}$, where K represent the number of groundtruth classes. The patches are sent to the proposed FusAtNet model which get processed as they pass through various modules that are discussed ahead in section 3.2.

3.1. Model overview

The intent behind this research work is to synergistically explore the specral-spatial properties of HSI and spatial/elevation characteristics of LiDAR modality using the "cross-attention" framework. The work of the attention modules is to selectively highlight the hotspots in the extracted hyperspectral features in order to increase the interclass variance and thus improve the classification accuracy. This is achieved in two steps: firstly, the HSI features are passed through a feature extractor and spectral attention module, and their combination is used to emphasize the spectral information in the HSI features. Simultaneously, the LiDAR features are passed through a spatial attention framework and the resultant mask accentuates the spatial characteristics of HSI. Secondly, the highlighted features are reinforced with the original features and passed through modality extraction and modality attention modules, the outputs of which are combined to judiciously highlight the important sections of the two modalities. The resultant features are then sent to the classification module.

3.2. Network architecture

The comprehensive architecture of FusAtNet is displayed in Fig. 3 and all the experiments adhere to the same. FusAtNet essentially contains six modules that are used in three phases. In the first phase, hyperspectral feature extractor $\mathcal{F}_{\mathcal{HS}}$, spectral attention module $\mathcal{A}_{\mathcal{S}}$ and spatial attention module $\mathcal{A}_{\mathcal{T}}$ are used to jointly extract and highlight the spatial-spectral features from the HSI. In the second phase, modality feature extractor $\mathcal{F}_{\mathcal{M}}$ and

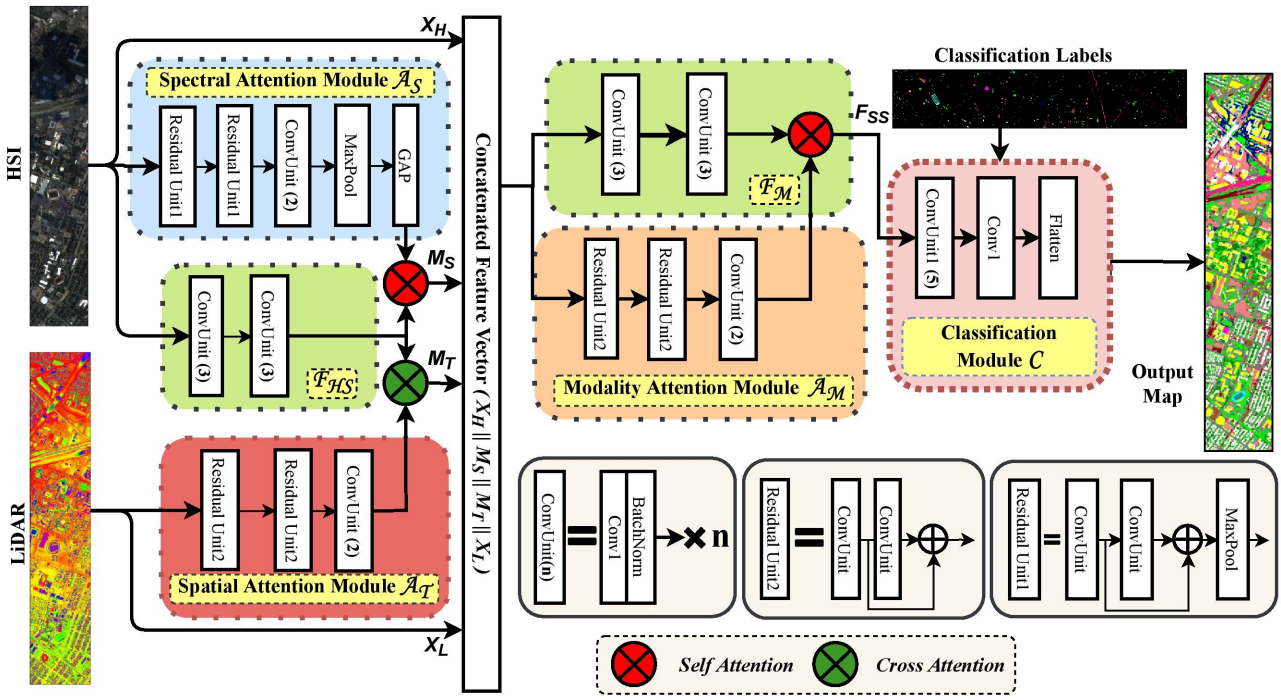

Figure 3. Schematic of FusAtNet (presented on Houston dataset). Initially, the hyperspectral training samples X_H are sent to the feature extractor $\mathcal{F}_{\mathcal{HS}}$ to get latent representations and to spectral attention module $\mathcal{A}_\mathcal{S}$ to generate spectral attention mask. Simultaneously, the corresponding LiDAR training samples X_L are sent to spatial attention module $\mathcal{A}_\mathcal{T}$ to get the spatial attention mask. The attention masks are individually multiplied to the latent HSI representations to get M_S and M_T. M_S and M_T are then concatenated with X_H and X_L and sent to modality feature extractor $\mathcal{F}_\mathcal{M}$ and modality attention module $\mathcal{A}_\mathcal{M}$. The outputs from the two are then multiplied to get F_{SS}, which is then sent to the classification module $\mathcal{C}$ for pixel classification.

modality attention module $\mathcal{A}_\mathcal{M}$ are used to selectively highlight the modality specific features. In the third phase, the modality specific spectral-spatial features are sent to the classification module $\mathcal{C}$. All the modules are inherently CNN modules where the size of all the kernels is fixed to 3×3 and non-linearity is fixed to *ReLU*. The modules are discussed as follows:

Hyperspectral feature extractor $\mathcal{F}_{\mathcal{HS}}$: $\mathcal{F}_{\mathcal{HS}}$ consists of a 6 layer CNN and is used to extract the spectral-spatial features from the HSIs. The first five layers contain 256 filters while the sixth layer has 1024 number of filters. All the convolution operations are applied with zero padding. Output of each convolution operation is operated on by batch normalisation. The module can be represented as $\mathcal{F}_{\mathcal{HS}}(\theta_{F_{HS}}, \mathbf{x}_H^i)$ where, θ_F represent the weights of the module. The output of $\mathcal{F}_{\mathcal{HS}}$ is a patch of size 11×11×1024.

Spectral attention module $\mathcal{A}_\mathcal{S}$: $\mathcal{A}_\mathcal{S}$ draws its attention mask from the HSI. The module is a CNN with 3 convolution blocks, with 2 convolution layers each. In addition, first and second convolution block are followed by a residual block each. There is a maxpooling layer after each residual block and the sixth convolution layer. The last layer of this module is a global average pooling (GAP) layer. Overall, the architecture of this model is inspired from [19]. The number of kernels in first five convolution layers is 256 and that in the sixth one is 1024, all of which use zero padding. Each convolution operation is followed by a batch normalisation layer. The model is denoted as $\mathcal{A}_\mathcal{S}(\theta_{A_S}, \mathbf{x}_H^i)$, where θ_{A_S} are the weights of this attention module. The output of this module is a vector of size 1×1024, which is multiplied with the output of $\mathcal{F}_{\mathcal{HS}}$ to get the highlighted spectral features as denoted in Eq. (1).

$$M_S(\mathbf{x}^i) = \mathcal{F}_{\mathcal{HS}}(\theta_{F_{HS}}, \mathbf{x}_H^i) \otimes \mathcal{A}_\mathcal{S}(\theta_{A_S}, \mathbf{x}_H^i) \qquad (1)$$

Here, M_S denotes the extracted features highlighted with spectral attention mask and $\otimes$ represent the broadcasted element-wise matrix multiplication operation (such that the resultant product retains the size of the matrix with higher dimension).

Spatial attention module $\mathcal{A}_\mathcal{T}$: Denoted by $\mathcal{A}_\mathcal{T}(\theta_{A_T}, \mathbf{x}_L^i)$, where θ_{A_T} denotes the weights, spatial attention module is a 6-layer CNN that generates attention mask from the LiDAR modality. The first 3 layers consist of 128 filters each while each of the last three layers has 256 number of filters. There are two residual layers, each after second and fourth convolution layer. All the convolution layers are followed by a batch normalisation operation. The output from

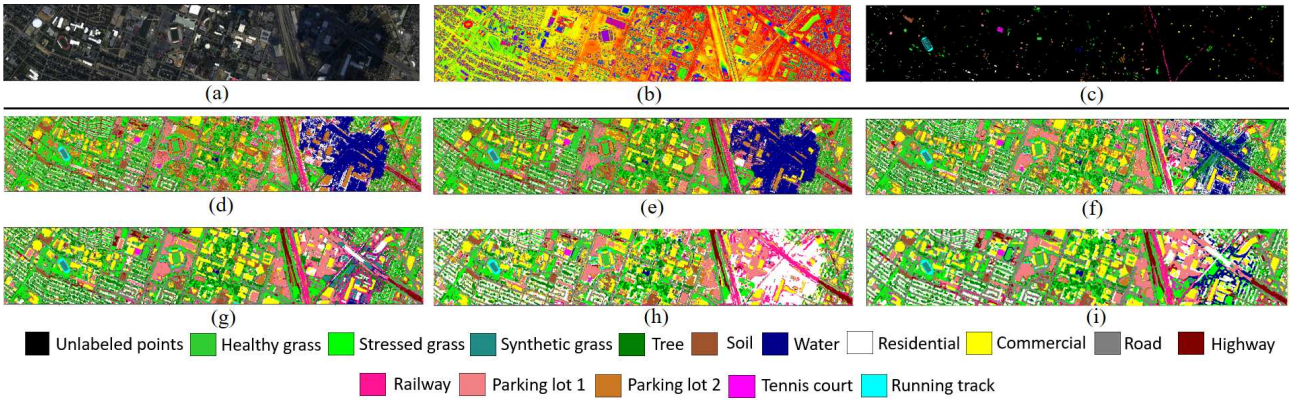

Figure 4. Houston hyperspectral and lidar dataset with classification maps. (a) True colour composite of HSI, (b) LiDAR image, (c) Groundtruth. Classification maps from (d) SVM (H), (e) SVM (H+L), (f) Two-branch CNN (H), (g) Two-branch CNN (H+L), (h) FusAtNet (H), (i) FusAtNet (H+L).

this module is a patch of size $11\times11\times1024$ that is multiplied with the extracted features from $\mathcal{F}_{\mathcal{HS}}$ to get spatially highlighted features M_T denoted in Eq. (2) as:

$$M_T(\mathbf{x}_H^i, \mathbf{x}_L^i) = \mathcal{F}_{\mathcal{HS}}(\theta_{F_{HS}}, \mathbf{x}_H^i) \otimes \mathcal{A}_{\mathcal{T}}(\theta_{A_T}, \mathbf{x}_L^i) \quad (2)$$

Modality feature extractor $\mathcal{F}_{\mathcal{M}}$: $\mathcal{F}_{\mathcal{M}}$ module follows the same structure as that of $\mathcal{F}_{\mathcal{HS}}$ and can be represented as $\mathcal{F}_{\mathcal{M}}(\theta_{F_M}, \mathbf{x}_H^i, \mathbf{x}_L^i)$, where θ_{F_M} are the weight of the module. It is fed with the spectrally and spatially highlighted features M_S and M_T along with the original $\mathcal{X}$, the output of which is a patch of size $11\times11\times1024$ which can be represented as in Eq. (3).

$$F_M(\mathbf{x}_H^i, \mathbf{x}_L^i) = \mathcal{F}_{\mathcal{M}}(\theta_{F_M}, \mathbf{x}_H^i \oplus \mathbf{x}_L^i \oplus M_S(\mathbf{x}^i) \oplus M_T(\mathbf{x}_H^i, \mathbf{x}_L^i)) \quad (3)$$

where, $\oplus$ represents concatenation along channel axis.

Modality attention module $\mathcal{A}_{\mathcal{M}}$: Architecture of $\mathcal{A}_{\mathcal{M}}$ is similar to that of $\mathcal{A}_{\mathcal{T}}$ and is denoted by $\mathcal{A}_{\mathcal{M}}(\theta_{A_M}, \mathbf{x}_H^i, \mathbf{x}_L^i)$, θ_{A_M} being the weights. The work of this module is to create an attention mask that focuses on specific traits of each modality and therefore the input is kept the same as that of $\mathcal{F}_{\mathcal{M}}$. This is represented in Eq. (4).

$$A_M(\mathbf{x}_H^i, \mathbf{x}_L^i) = \mathcal{A}_{\mathcal{M}}(\theta_{A_M}, \mathbf{x}_H^i \oplus \mathbf{x}_L^i \oplus M_S(\mathbf{x}^i) \oplus M_T(\mathbf{x}_H^i, \mathbf{x}_L^i)) \quad (4)$$

The output of the module is an $11\times11\times1024$ patch that is multiplied with the output of $\mathcal{F}_{\mathcal{M}}$, as shown in Eq. (5), and the result is sent to the classification module.

$$F_{SS}(\mathbf{x}_H^i, \mathbf{x}_L^i) = F_M(\mathbf{x}_H^i, \mathbf{x}_L^i) \otimes A_M(\mathbf{x}_H^i, \mathbf{x}_L^i) \quad (5)$$

where, F_{SS} are the final spectral-spatial features.

Classification module $\mathcal{C}$: The input to $\mathcal{C}$ module are the final spectral-spatial features $F_{SS}(\mathbf{x}_H^i, \mathbf{x}_L^i)$. The module is a 6-layer fully convolutional neural network where first four layers consist of 256 filters each while the fifth and sixth layers respectively contain 1024 and K filters, where K is the number of classes. The filter size for the last layer is set to 1×1 and no padding is used in any layer. All the layers except last one are operated on by *ReLU* activation function and batch normalisation while the last layer is the softmax layer. The module can be defined as $\mathcal{C}(\theta_C, F_{SS}(\mathbf{x}_H^i, \mathbf{x}_L^i))$ where, θ_C are classification weights. The output of $\mathcal{C}$ is a vector of size $1\times$K.

3.3. Training and inference

The output from $\mathcal{C}$ is subjected to a categorical cross-entropy loss which is backpropagated to train the FusAtNet model in an end-to-end fashion (refer Eq. (6)).

$$\mathcal{L}_C = -\mathbb{E}_{(\mathbf{x}_H^i, \mathbf{x}_L^i, y^i)}[y^i \log \mathcal{C}(\theta_C, F_{SS}(\mathbf{x}_H^i, \mathbf{x}_L^i))] \quad (6)$$

where, $\mathcal{L}_C$ is the classification loss.

During the testing phase, the given test sample $(\mathbf{x}_H^j, \mathbf{x}_L^j)$ is passed through the fusion module and follows the same path as that of the training samples. The resultant output $F_{SS}(\mathbf{x}_H^j, \mathbf{x}_L^j)$ is sent to the classification module $\mathcal{C}$ where it is assigned the predicted class label.

4. Experimental setup

This section discusses about the datasets used to validate FusAtNet and protocols followed while training the same.

4.1. Datasets

To evaluate the efficacy our method, three HSI-LiDAR datasets have been considered.

Houston dataset: This dataset consists of a hyperspectral imagery and a LiDAR depth raster and was introduced in GRSS Data Fusion Contest 2013. The dataset is acquired over the Houston university campus and surroundings by

National Airborne Centre for laser mapping (NCALM). The HSI is composed of 144 hyperspectral bands with the wavelengths varying from 0.38 μm to 1.05 μm with each raster of size 349×1905 and spatial resolution 2.5 m. A total of 15029 groundtruth samples are available that are distributed over 15 classes and divided into training and testing sets containing 2832 and 12197 pixels respectively [40]. However, for our experiments, 12189 pixels are considered in the test set since a few of the pixels were interfering with the data preprocessing. The dataset can be visualised in Fig. 4.

Trento dataset: This dataset is collected using AISA eagle sensor over the rural regions in Trento, Italy. The HSI

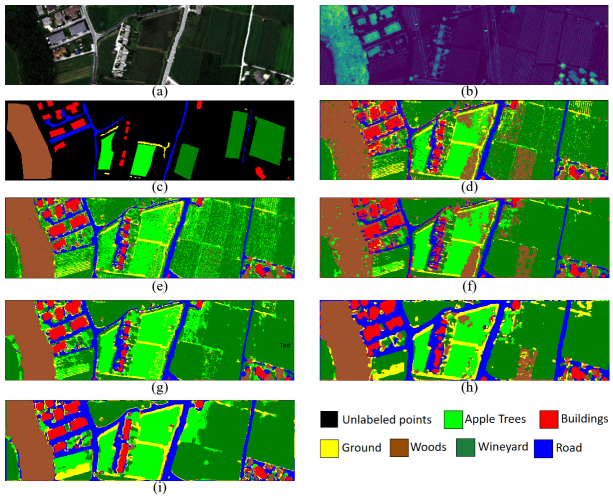

Figure 5. Trento hyperspectral and lidar dataset with classification maps. (a) True colour composite of HSI, (b) LiDAR image, (c) Groundtruth. Classification maps from (d) SVM (H), (e) SVM (H+L), (f) Two-branch CNN (H), (g) Two-branch CNN (H+L), (h) FusAtNet (H), (i) FusAtNet (H+L).

image is composed of 63 bands with their wavelengths in the range of 0.42 μm to 0.99 μm, while LiDAR consists of 2 rasters showing elevation data. The dimension of each band is 166 × 600 while the spatial and spectral resolutions are 9.2 nm and 1.0 m respectively. There are a total of 6 classes in the imagery, groudtruth of which are available for 30214 pixels that are divided in 819 training pixels and 29395 test pixels [40]. The dataset is displayed in Fig. 5.

MUUFL Gulfport dataset: This dataset is acquired over the campus of University of Southern Mississippi Gulf Park, Long Beach Mississippi in November, 2010. The HSI imagery originally contained 72 bands. However, due to noise, initial and final four bands are omitted leading to a total of 64 bands. The LiDAR modality consists of two elevation rasters. All the bands and rasters are coregistered, acquiring the total size of 325×220. There are a total of 53687 groundtruth pixels encompassing 11 classes [41, 42]. For training, 100 pixels per class are selected leaving the total of 52587 pixels for testing. The HSI and LiDAR imageries along with the groundtruth pixels can be viewed in Fig. 6.

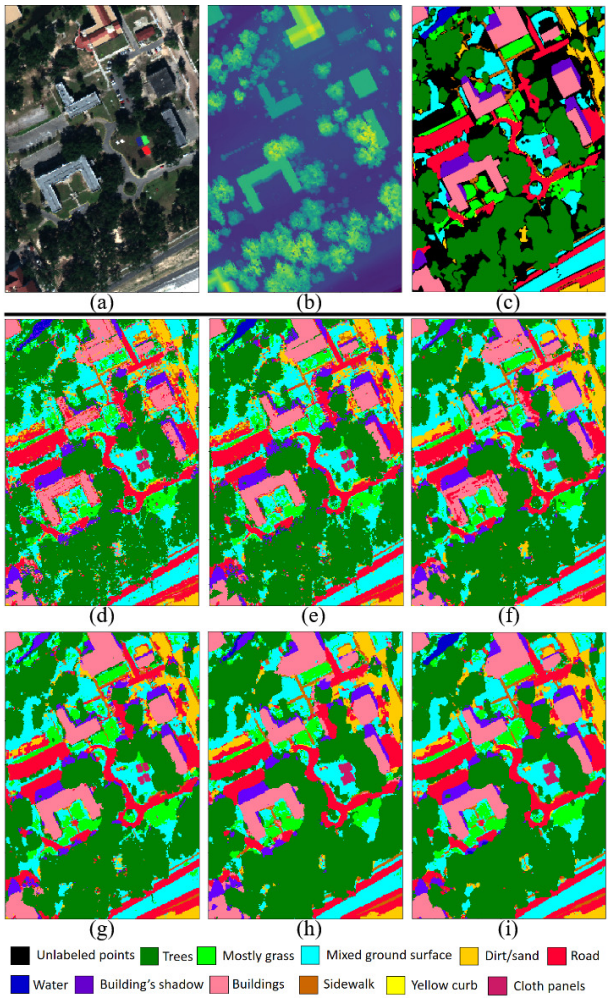

Figure 6. MUUFL hyperspectral and lidar dataset with classification maps. (a) True colour composite of HSI, (b) LiDAR image, (c) Groundtruth. Classification maps from (d) SVM (H), (e) SVM (H+L), (f) Two-branch CNN (H), (g) Two-branch CNN (H+L), (h) FusAtNet (H), (i) FusAtNet (H+L).

4.2. Training protocols

Our method is compared against other conventional and state of the art multimodal learning methods from [40] to fuse HSI and LiDAR modalities, such as SVM [43], extreme learning machines [44], CNN-PPF [45] and two-branch CNN [40] with spectral and spatial feature extraction. The SVM (both hyperspectral and LiDAR) and ELM (only hyperspectral) models for Trento dataset have been retrained and re-evaluated since the values in [40] seemed incorrect. All the analyses have been carried out on both HSI only (represented as (H) in the results and classified maps) as well fused HSI and LiDAR (represented as (H+L)) data

Table 1. Accuracy analysis on the Houston dataset (in %). 'H' represents only HSI while 'H+L' represents fused HSI and LiDAR.

Classes	SVM (H) [43]	SVM (H+L) [43]	ELM (H) [44]	ELM (H+L) [44]	CNN-PPF (H) [45]	CNN-PPF (H+L) [45]	Two Branch CNN (H) [46]	Two Branch CNN (H+L) [46]	Proposed (H)	Proposed (H+L)
Healthy Grass	81.86	82.43	82.91	83.10	82.24	**83.57**	83.38	83.10	83.00	83.10
Stressed Grass	82.61	82.05	83.93	83.70	**98.31**	98.21	84.21	84.10	84.96	96.05
Synthetic Grass	99.80	99.80	**100.00**	**100.00**	70.69	98.42	99.60	**100.00**	**100.00**	**100.00**
Trees	92.50	92.80	91.76	91.86	94.98	**97.73**	93.18	93.09	92.23	93.09
Soil	98.39	98.48	98.77	98.86	97.25	96.50	98.58	**100.00**	97.06	99.43
Water	94.41	95.10	95.10	95.10	79.02	97.20	99.30	99.30	**100.00**	**100.00**
Residential	76.87	75.47	89.65	80.04	86.19	85.82	85.45	92.82	**93.81**	93.53
Commercial	43.02	46.91	49.76	68.47	65.81	56.51	69.14	82.34	76.35	**92.12**
Road	79.04	77.53	81.11	84.80	72.11	71.20	78.66	84.70	**85.15**	83.63
Highway	58.01	60.04	54.34	49.13	55.21	57.21	52.90	**65.44**	62.64	64.09
Railway	81.59	81.02	74.67	80.27	85.01	80.55	82.16	88.24	72.11	**90.13**
Parking Lot 1	72.91	85.49	69.07	79.06	60.23	62.82	**92.51**	89.53	88.95	91.93
Parking Lot 2	71.23	75.09	69.82	71.58	75.09	63.86	92.63	92.28	**92.98**	88.42
Tennis Court	99.60	**100.00**	99.19	99.60	83.00	**100.00**	94.33	96.76	**100.00**	**100.00**
Running Track	97.67	98.31	98.52	98.52	52.64	98.10	99.79	99.79	**100.00**	99.15
OA	79.00	80.49	79.87	81.92	78.35	83.33	84.08	87.98	85.72	**89.98**
AA	81.94	83.37	82.57	84.27	77.19	83.21	86.98	90.11	88.62	**94.65**
κ	0.7741	0.7898	0.7821	0.8045	0.7646	0.8188	0.8274	0.8698	0.8450	**0.8913**

Table 2. Accuracy analysis on the Trento dataset (in %). 'H' represents only HSI while 'H+L' represents fused HSI and LiDAR.

Classes	SVM (H) [43]	SVM (H+L) [43]	ELM (H) [44]	ELM (H+L) [44]	CNN-PPF (H) [45]	CNN-PPF (H+L) [45]	Two Branch CNN (H) [40]	Two Branch CNN (H+L) [40]	Proposed (H)	Proposed (H+L)
Apples	90.80	85.49	91.32	95.81	92.22	95.88	98.04	98.07	99.06	**99.54**
Buildings	84.22	89.76	85.74	96.97	87.08	**99.07**	97.45	95.21	97.05	98.49
Ground	98.12	59.56	97.59	96.66	66.81	91.44	83.09	93.32	**100.00**	99.73
Woods	97.01	97.42	88.44	99.39	65.24	99.79	98.29	99.93	**100.00**	**100.00**
Vineyard	79.02	93.85	86.39	82.24	98.98	98.56	98.29	98.78	99.85	**99.90**
Roads	66.92	89.96	64.06	86.52	73.19	88.72	68.21	89.98	89.39	**93.32**
OA	85.56	92.30	85.43	91.32	83.52	97.48	95.35	97.92	98.50	**99.06**
AA	86.02	86.01	85.59	92.93	80.59	95.58	90.86	96.19	97.56	**98.50**
κ	0.8102	0.8971	0.8065	0.9042	0.7843	0.9664	0.9379	0.9681	0.9796	**0.9875**

Table 3. Accuracy analysis on the MUUFL dataset (in %). 'H' represents only HSI while 'H+L' represents fused HSI and LiDAR.

Classes	SVM (H) [43]	SVM (H+L) [43]	ELM (H) [44]	ELM (H+L) [44]	Two Branch CNN (H) [40]	Two Branch CNN (H+L) [40]	Proposed (H)	Proposed (H+L)
Trees	93.91	95.97	91.99	94.89	97.07	97.40	97.74	**98.10**
Grass Pure	59.54	62.71	39.44	62.23	62.93	**76.84**	63.71	71.66
Grass Groundsurface	82.72	83.60	76.87	83.15	87.44	84.31	86.48	**87.65**
Dirt and Sand	79.11	78.60	74.51	57.88	**90.74**	84.93	87.34	86.42
Road Materials	91.20	92.72	92.14	93.33	85.30	93.41	93.07	**95.09**
Water	54.31	95.10	0.00	68.32	5.39	10.78	24.78	**90.73**
Buildings' Shadow	58.35	71.23	63.88	47.01	67.33	63.34	72.55	**74.27**
Buildings	75.94	87.96	68.26	77.58	82.33	96.20	96.38	**97.55**
Sidewalk	43.85	41.11	24.22	32.15	54.59	54.30	56.07	**60.44**
Yellow Curb	11.05	11.05	0.00	0.00	**24.31**	2.21	7.73	9.39
Cloth Panels	88.37	88.76	89.92	78.29	90.31	87.21	92.25	**93.02**
OA	83.39	86.90	78.49	83.10	86.30	89.38	89.41	**91.48**
AA	67.12	83.37	56.48	63.17	67.97	68.26	70.74	**78.58**
κ	0.7790	0.8255	0.7137	0.7742	0.8197	0.8583	0.8581	**0.8865**

to affirm the efficacy of multimodal learning over unimodal learning. To assess the performance of the methods, overall accuracy (OA), producer's accuracy (PA), average accuracy (AA) and Cohen's kappa (κ) have been used as evaluation metrics. Both HSI and LiDAR data is subjected to min-max normalisation to scale the modalities and speed up the convergence.

The network uses a fixed patch size of 11×11 for all the datasets. These patches are created around the pixel with known groundtruth label. In addition, in order to boost the performance of our model, we resort to data augmentation technique (used in [5]) by rotating the training patches by 90°, 180° and 270° in clockwise direction. All the weight initialization are carried out using glorot initialization[47] while the training is performed for 1000 epochs. A small initial learning rate of 0.000005 is chosen because a higher learning rate leads to higher fluctuations when Adam optimizer is used with Nesterov momentum [48].

5. Results and discussion

Our proposed method is verified on Houston, Trento and MUUFL datasets in Tables 1, 2 and 3 respectively. It is clearly visible for all the cases that our method outperforms all the state of the art methods with a significant margin in all the avenues, be it OA (the respective accuracies of Houston, Trento and MUUFL datasets being 89.98%, 99.06% and 91.48%), AA (respective values being 94.65%, 98.50% and 78.58%) or κ. It is also easily observed that

in case of classwise/producer's accuracy, the performance of our method is better than the other methods for most of the classes and only marginally exceeded by other methods for a few of them. For Houston dataset, it can be noted that the accuracy for 'commercial' class (92.12%) is significantly improved for our method in comparison to other methods. This can be attributed to the fact that commercial regions generally have a variable layout with frequent elevation changes that are effectively captured by LiDAR based attention maps. Similarly, in case of Trento dataset, the 'road' class shows a notable increase in accuracy (93.32%). This increment is also on account of variation in road profile with respect to its elevation. The classification maps for Houston, Trento and MUUFL datasets are presented in Fig. 4, 5 and 6 respectively. It can be visually verified that the classification maps obtained from FusAtNet tend to be less noisy and have smooth interclass transitions. It is also observed in Fig. 4 that methods such as SVM and two-branch CNN tend to classify the shadowy areas as water (in the right portion of the maps) because of their darker tone. Our approach largely mitigates this problem as well.

5.1. Ablation study

We further carried out different ablation studies to highlight the individual aspects of our model. In table 4, we evaluate our model's performance by iteratively removing each of the attention module. It is evidently visible that in absence of even anyone of the attention module, the model tends to underperform. In addition, the importance of spatial characteristics of LiDAR modality is also proven since the presence of only LiDAR based spatial attention module gives better accuracy than HSI based spectral attention module for all the three datasets.

Table 4. Ablation study by changing attention layers on all the datasets (accuracy in %).

Attention layers	OA (Houston)	OA (Trento)	OA (MUUFL)
Only $\mathcal{A}_{\mathcal{S}}$	86.48	98.04	90.09
Only $\mathcal{A}_{\mathcal{T}}$	88.39	98.79	91.13
Only $\mathcal{A}_{\mathcal{M}}$	87.51	98.48	89.31
Only $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{T}}$	89.04	98.77	90.69
Only $\mathcal{A}_{\mathcal{T}}$ and $\mathcal{A}_{\mathcal{M}}$	87.78	98.95	98.24
Only $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{M}}$	86.90	98.24	89.14
All $\mathcal{A}_{\mathcal{S}}$, $\mathcal{A}_{\mathcal{T}}$ and $\mathcal{A}_{\mathcal{M}}$	**89.98**	**99.06**	**91.48**

Table 5 displays the performance of our method when trained without data augmentation. Since our model is quite deep, there is a decrease in the performance when no augmentation is applied on training samples. This magnitude of this decrease is maximum in case of Houston dataset (4.76%) since it has most number of features in comparison to other datasets. Hence, it requires comparatively more iterations to converge and give better accuracy.

Furthermore, an additional ablation study is carried out on all the datasets to check the effect of decreasing the training size and then evaluate the performance of our model as displayed in table 6. As expected, the accuracy progressively decreases as the number of training samples decrease, further reinforcing the high data requirement of the deep learning models.

Table 5. Ablation study for training with and without data augmentation (accuracy in %).

Data	OA (Houston)	OA (Trento)	OA (MUUFL)
No augmentation	85.22	98.32	88.81
With augmentation	**89.98**	**99.06**	**91.48**

Table 6. Model performance by changing the fraction of training samples on MUUFL dataset (accuracy in %).

	30% data	50% data	75% data	100% data
OA (Houston)	84.75	87.92	88.66	**89.98**
OA (Trento)	97.78	98.48	98.93	**99.06**
OA (MUUFL)	86.90	89.58	89.78	**91.48**

6. Conclusions and future work

We introduce a novel fusion network for HSI and LiDAR data for the purpose of producing improved land-cover maps. Our network, called FusAtnet, judiciously utilizes different attention learning modules to learn joint feature representations given both the input modalities. To this end, we propose the notion of cross-attention where the feature learning stream for a given modality is influenced by the other modality. The results obtained for multiple datasets confirm the efficacy of the proposed fusion network. Due to the generic nature of FusAtNet, it can be extended to support a varied range of modalities with minimum overhead. In future, we plan to extend the network to support more than two modalities. Besides, we also plan to perform rigorous model engineering to limit the number of learnable parameters without compromising the performance, for example, using the notion of dilated convolution in the attention modules effectively.

Acknowledgement

The authors would like to thank the reviewers for their comments that helped to improve the quality of this paper. B. Banerjee was supported by SERB, DST Grant ECR/2017/000365. S. Mohla acknowledges support from Shastri Indo-Canadian Institute through SRSF research fellowship.

References

[1] Benjamin Koetz, Felix Morsdorf, Sebastian Van der Linden, Thomas Curt, and Britta Allgöwer. Multi-source land cover classification for forest fire management based on imaging spectrometry and lidar data. *Forest Ecology and Management*, 256(3):263–271, 2008. 1

[2] Wei Li, Eric W Tramel, Saurabh Prasad, and James E Fowler. Nearest regularized subspace for hyperspectral classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):477–489, 2013. 1

[3] Lian-Zhi Huo, Carlos Alberto Silva, Carine Klauberg, Midhun Mohan, Li-Jun Zhao, Ping Tang, and Andrew Thomas Hudak. Supervised spatial classification of multispectral lidar data in urban areas. *PloS one*, 13(10), 2018. 1

[4] Christian Debes, Andreas Merentitis, Roel Heremans, Jürgen Hahn, Nikolaos Frangiadakis, Tim van Kasteren, Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Sidharta Gautama, et al. Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2405–2418, 2014. 1

[5] Mengmeng Zhang, Wei Li, Qian Du, Lianru Gao, and Bing Zhang. Feature extraction for classification of hyperspectral and lidar data using patch-to-patch cnn. *IEEE transactions on cybernetics*, 2018. 1, 7

[6] Hassan Ghassemian. A review of remote sensing image fusion methods. *Information Fusion*, 32:75–89, 2016. 1

[7] Grigorios Tsagkatakis, Anastasia Aidini, Konstantina Fotiadou, Michalis Giannopoulos, Anastasia Pentari, and Panagiotis Tsakalides. Survey of deep-learning approaches for remote sensing observation enhancement. *Sensors*, 19(18):3929, 2019. 1

[8] E Simental, DJ Ragsdale, E Bosch, R Dodge Jr, and R Pazak. Hyperspectral dimension reduction and elevation data for supervised image classification. In *Proc. 14th ASPRS Conf*, pages 3–9, 2003. 2

[9] Dirk Lemp and Uwe Weidner. Improvements of roof surface classification using hyperspectral and laser scanning data. In *Proc. ISPRS Joint Conf.: 3rd Int. Symp. Remote Sens. Data Fusion Over Urban Areas (URBAN), 5th Int. Symp. Remote Sens. Urban Areas (URS)*, pages 14–16, 2005. 2

[10] Michele Dalponte, Lorenzo Bruzzone, and Damiano Gianelle. Fusion of hyperspectral and lidar remote sensing data for classification of complex forest areas. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1416–1427, 2008. 2

[11] Behnood Rasti, Pedram Ghamisi, Javier Plaza, and Antonio Plaza. Fusion of hyperspectral and lidar data using sparse and low-rank component analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6354–6365, 2017. 2

[12] Junshi Xia, Zuheng Ming, and Akira Iwasaki. Multiple sources data fusion via deep forest. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1722–1725. IEEE, 2018. 2

[13] Maximilian Brell, Karl Segl, Luis Guanter, and Bodo Bookhagen. Hyperspectral and lidar intensity data fusion: A framework for the rigorous correction of illumination, anisotropic effects, and cross calibration. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2799–2810, 2017. 2

[14] Luis Gómez-Chova, Devis Tuia, Gabriele Moser, and Gustau Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, 2015. 2

[15] Dushyant Rao, Mark De Deuge, Navid Nourani-Vatani, Stefan B Williams, and Oscar Pizarro. Multimodal learning and inference from visual and remotely sensed data. *The International Journal of Robotics Research*, 36(1):24–43, 2017. 2

[16] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li. Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, 2015, 2015. 2

[17] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected topics in applied earth observations and remote sensing*, 7(6):2094–2107, 2014. 2

[18] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018. 2

[19] Lichao Mou and Xiao Xiang Zhu. Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2019. 2, 3, 4

[20] Juan Mario Haut, Mercedes E Paoletti, Javier Plaza, Antonio Plaza, and Jun Li. Visual attention-driven hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):8065–8080, 2019. 2, 3

[21] J Ngiam, A Khosla, M Kim, J Nam, H Lee, and A Ng. Multimodal deep learning,[in proc. 28th int. conf. *Mach. Learn*, 2011. 2

[22] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. Structure-aware multimodal feature fusion for rgb-d scene classification and beyond. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2s):1–22, 2018. 2

[23] Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, and Mihai Datcu. Cmir-net: A deep learning based model for cross-modal retrieval in remote sensing. *Pattern Recognition Letters*, 2020. 2

[24] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *NeuroImage*, 101:569–582, 2014. 3

[25] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169, 2019. 3

[26] Muxuan Liang, Zhizhong Li, Ting Chen, and Jianyang Zeng. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(4):928–937, 2014. 3

[27] Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Sidharta Gautama, and Wilfried Philips. Combining feature fusion and decision fusion for classification of hyperspectral and lidar data. In *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 1241–1244. IEEE, 2014. 3

[28] Chongyue Zhao, Xinbo Gao, Ying Wang, and Jie Li. Efficient multiple-feature learning-based hyperspectral image classification with limited training samples. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7):4052–4062, 2016. 3

[29] Mehrdad Eslami and Ali Mohammadzadeh. Developing a spectral-based strategy for urban object detection from airborne hyperspectral tir and visible data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(5):1808–1816, 2015. 3

[30] Yong Yang, Weiguo Wan, Shuying Huang, Feiniu Yuan, Shouyuan Yang, and Yue Que. Remote sensing image fusion based on adaptive ihs and multiscale guided filter. *IEEE Access*, 4:4573–4582, 2016. 3

[31] Joy Jinju, N Santhi, K Ramar, and B Sathya Bama. Spatial frequency discrete wavelet transform image fusion technique for remote sensing applications. *Engineering Science and Technology, an International Journal*, 22(3):715–726, 2019. 3

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[33] Saurabh Morchhale, V Paúl Pauca, Robert J Plemmons, and Todd C Torgersen. Classification of pixel-level fused hyperspectral and lidar data using deep convolutional neural networks. In *2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–5. IEEE, 2016. 3

[34] Yushi Chen, Chunyang Li, Pedram Ghamisi, Chunyu Shi, and Yanfeng Gu. Deep fusion of hyperspectral and lidar data for thematic classification. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3591–3594. IEEE, 2016. 3

[35] Yushi Chen, Chunyang Li, Pedram Ghamisi, Xiuping Jia, and Yanfeng Gu. Deep fusion of remote sensing data for accurate classification. *IEEE Geoscience and Remote Sensing Letters*, 14(8):1253–1257, 2017. 3

[36] Quanlong Feng, Dehai Zhu, Jianyu Yang, and Baoguo 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. 3

[37] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 3

[38] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 3

[39] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[40] Xiaodong Xu, Wei Li, Qiong Ran, Qian Du, Lianru Gao, and Bing Zhang. Multisource remote sensing data classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):937–949, 2017. 6, 7

[41] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell. MUUFL Gulfport Hyperspectral and LiDAR Airborne Data Set. Tech. Rep. REP-2013-570, University of Florida,, Gainesville, FL, October 2013. 6

[42] X. Du and A. Zare. Technical Report: Scene Label Ground Truth Map for MUUFL Gulfport Data Set. Tech. Rep. 20170417, University of Florida, Gainesville, FL, April 2017. 6

[43] Grégoire Mercier and Marc 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. 6, 7

[44] Wei Li, Chen Chen, Hongjun Su, and Qian Du. Local binary patterns and extreme learning machine for hyperspectral imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3681–3693, 2015. 6, 7

[45] Wei Li, Guodong Wu, Fan Zhang, and Qian Du. Hyperspectral image classification using deep pixel-pair features. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):844–853, 2016. 6, 7

[46] Tao Xu, Han Zhang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas. Multimodal deep learning for cervical dysplasia diagnosis. In *International conference on medical image computing and computer-assisted intervention*, pages 115–123. Springer, 2016. 7

[47] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 7

[48] Timothy Dozat. Incorporating nesterov momentum into adam. 2016. Accessed: December 8, 2019 [Online]. Available: http://cs229.stanford.edu /proj2015/054_report.pdf. 7