Pansharpening and Semantic Segmentation of Satellite Imagery

Nishchal J
Computer Science
RV College of Engineering
Bangalore, India
nishchalj.cs18@rvce.edu.in

Sanjana Reddy
Computer Science
RV College of Engineering
Bangalore, India
sanjanasr.cs18@rvce.edu.in

Navya Priya N
Computer Science
RV College of Engineering
Bangalore, India
navyapriyan.cs18@rvce.edu.in

Varsha R Jenni
Computer Science
RV College of Engineering
Bangalore, India
varsharjenni.cs18@rvce.edu.in

Hebbar R
RRSC-South
NRSC/ISRO
Bangalore, India
hebbar_kr@nrsc.gov.in

B. Sathish Babu
Computer Science
RV College of Engineering
Bangalore, India
bsbabu@rvce.edu.in

***Abstract*— The fusion of high spatial resolution panchromatic (PAN) data with simultaneously acquired lower spatial resolution multispectral (MS) data is called pansharpening. The conventional techniques like Brovey and IHS to perform the same are lengthy and time consuming. Hence, this paper proposes a Convolutional Neural Network for pansharpening which gives better results than the conventional methods. This paper also looks at binary and multiclass semantic segmentation using U-Net and its variations. Finally, a model based on 3D convolutions is introduced which performs semantic segmentation on Hyperspectral Imagery with high accuracy and computational efficiency.**

Keywords—Multispectral imagery, panchromatic images, pansharpening, PansharpNet, Convolutional Neural Network (CNN), Swish activation, semantic segmentation, U-Net, Hyperspectral imagery, 3D-Hyper UNET, 3D Convolutions

I. Introduction

Satellite images are one of the most powerful and important tools used in the fields of farming, weather forecasting, disaster mitigation planning and recovery etc. They have enormous potential for qualitative estimation of land and water cover on earth. Due to lack of quality the remote sensed images cannot be used directly for observations. The good quality images have to be preprocessed, classified or segmented inorder to gain useful instincts.

A high-resolution multispectral image, i.e., with high spatial and spectral resolution, cannot be obtained directly from the remote sensors. Hence, techniques like pansharpening are used to improve spatial and spectral resolution, where a high spatial resolution panchromatic (PAN) component with a low spectral resolution, and a multispectral (MS) component with complementary properties are fused. Multispectral images have bands between 3 and 16 (RGB + NIR) and panchromatic images have a single grey band. These images can be used individually to get enough information but, full advantage of the available information can be taken when these two images are preprocessed and fused through various pansharpening techniques. Pansharpening is useful in interpreting a remote sensing scene, segmentation, classification and feature extraction by utilising bands which are not visible to the naked eye. Artificial intelligence further simplifies this process as conventional techniques are rather cumbersome and time consuming.

Semantic segmentation of satellite imagery is popularly implemented to recognize which class each pixel of an image belongs to. A CNN would generally be used to pin down which class a given satellite image belongs to. But in numerous visual projects, a single class as output for an entire image is not sufficient. There is a need to classify every pixel, to generate a segmented image. The segmentation can be binary, to focus on a particular type of land cover like water bodies or vegetation, or multiclass to segment the image into various classes. Building a CNN for this purpose needs a gigantic amount of segmented ground truth masks, which is tedious to produce. Master annotators may be required for accurately distinguishing pixels situated at object limits and vague locales [12].

Hyperspectral Imagery(HSI) consists of a continuous spectrum of bands in the infrared, visible and other domains of the electromagnetic spectrum. Each pixel in this data will have three dimensions - two spatial and a deep spectral dimension. The number of spectral bands involved range from 50 to 300 or more depending on the number of wavelengths used to generate the image. Hyperspectral image analysis(HIA) is gaining scope due to its newly discovered applications in agriculture, industry and surveillance. With respect to classification and segmentation, most of the techniques use the classification and segmentation solutions meant for RGB images. Some techniques reduce the bands significantly to a lower number and carry out the classification procedures. The main disadvantage of these techniques is the enormous spectral information provided by HSI is simply ignored, which could otherwise aid to provide better classification and segmentation results.

In this paper, a novel pansharpening CNN is proposed for pansharpening to obtain high resolution multispectral data. It utilises the complementary resolutions and bands of both panchromatic and multispectral images to learn and predict results. Since the satellite images have high spatial resolution, they are broken down into smaller chunks and built back once the desired features are extracted and fused. The pansharpened images can then be used to carry out the segmentation tasks which are also explored in this paper. Finally, this paper also proposes a novel CNN called the 3-D Hyper-UNET for hyperspectral image segmentation. It utilises the spatial and spectral features of HSI using 3-D convolutions to analyse and provide better results. Further, the 3-D convolution transpose layers are used to build the segmentation mask from the feature vector extracted through the 3D convolution process.

II. Literature Survey

Many conventional algorithms like simple brovey, simple mean, ESRI etc are used to enhance the spectral and spatial resolution of satellite imagery. By using the influence of weights in IHS and Brovey methods which are conventional algorithms, an attempt was made for pansharpening using the WorldView-3 satellite images by Parente C [1]. The impact of weights which was obtained by spectral radiance response was being utilised for pansharpening and indexes like root mean square error, relative average spectral error were used to grade the quality of the outputs. Introducing different weights for each band of WV-3 images, IHS and Brovey methods gave better results, however no single pansharpening method could be considered the best as different methods performed well on different datasets.

In the work presented by Jaewan Choi, an attempt was made to perform pansharpening using guided filtering(GF) [2] on very high resolution satellite images. Due to contrast in their spectral and spatial characteristics and delays between panchromatic and multispectral sensors, the pansharpening process was characterized by spatial dissimilarities. The experimental results showed that the proposed method yields less spectral distortion and better spatial clarity than conventional pansharpening algorithms. However, since GF required a relatively high computational cost, further work using parallel processing or graphics processing units was needed.

In the work presented by Giuseppe Masi[3], a CNN was proposed for pansharpening which used MS components and index maps by upsampling and joining with the PAN at the input stage. This technique is called interpolation. However, all the features of multispectral images weren't extracted but were merely concatenated with panchromatic images, hence full advantage of the available information wasn't taken into account. Work has been done using GAN on image fusion by Zhiguang Yang[4] where a novel multi-exposure (over exposure and under exposure) image fusion method based on generative adversarial networks (termed as GANFuse) was introduced. It utilized GAN to fuse infrared and RGB images which achieved good performance. However for moving objects in image, this method failed to extract features and caused the ghost effect.

As there are no dedicated models for pansharpening, this paper has proposed a Pansharpening CNN architecture which is an improvement to [3], by considering all the bands and features available for pansharpening.

Semantic segmentation for vegetation cover detection in satellite images is a new field of interest. It has been seen that the use of U-Net for semantic segmentation of biomedical images, even with a smaller dataset, gives good results[6]. This work heavily relies on data augmentation with elastic deformation to ensure the proper use of the smaller dataset. There have been efforts to detect buildings from satellite images with low resolution, using models inspired from U-Net [7]. A method to handle the less accurate segmentation along the edges is also proposed. It involves weighing the inner and outer boundaries of the buildings. A revised U-Net architecture is proposed which has a pre-trained ResNet50 model as an encoder. The revised U-Net model displayed a good amount of improvement on the low resolution dataset that was available [8]. There have been trials of different Fully Convoluted Network (FCN) architectures for the same purpose. FCN architectures with VGG, GoogleNet and ResNet as base networks are discussed for semantic segmentation. A possibility of adding a discriminator network is also proposed. This creates an adversarial setting for the model[9].

Classification and segmentation of RGB data has been carried out for a long time and various CNN's have been implemented on datasets like MNIST, CIFAR, IMAGENET etc.. Qishuo Gao[10] had proposed a model for HSI classification implemented on Indian Pines, Univ. of Pavia, and Salinas Scene data. However this method used the normal 2D CNN's for classification and no segmentation was involved. Swalpa Kumar Roy[11] proposed an architecture called SNHybrid involving a mixture of 2D and 3D CNN for HSI Classification on the same datasets as the previous. This model's accuracy significantly showed improvement when compared to its predecessors, however, this too had no segmentation involved. Thus, the model proposed in this paper is an improvement of the SNHybrid architecture which performs segmentation of Hyperspectral Data using 3-D CNN's for feature extraction and segmentation.

III. Proposed Pansharpnet model

The whole network is built in three blocks, with the first block for multispectral image and second block for panchromatic image. The output of these blocks are concatenated to give the pansharpened image in the output block. The model takes in two inputs- a multispectral image (size (X',Y',Z) where X' * Y' represents the spatial resolution and Z represents the spectral bands) and panchromatic image (of size (X,Y,1) where X*Y represents the spatial resolution). It is understood that X*Y > X'*Y' and Z>=3.

For the multispectral image in block 1, a convolution layer of 16 filters with filter size (2,2) is used followed by a max pooling layer of stride (3,3). This combination of convolution and maxpool layers are repeated thrice with 32, 64 and 128 filters. This block is thus responsible for extraction of features from the low resolution multispectral images.

For the panchromatic image in block 2, a convolutional layer with 16 filters of size (3,3) is used followed by a max pooling

layer of stride (2,2). This combination of conv and maxpool layers are repeated thrice with 32, 64 and 128 filters.

Finally the convolution layer's output from block 1 and max pooling layer's output from block 2 are concatenated and transposed convolutionally to upscale the image. Further, the transposed output is concatenated with the corresponding features of block 1and 2 as shown in the FIG 1. Finally , the last layer is a convolutional layer of 4 filters with stride (1,1) producing the high resolution hyperspectral image.

A. Experiments and discussion

1) Dataset Description

Bangalore Urban Satellite Imagery data has been used for the training, validation and testing process. The pansharpened image - ground truth has been obtained from the software QGIS using the multispectral and panchromatic image of Bangalore satellite image. QGIS is an open-source cross platform geographic information system application used which supports viewing, editing and analysis of geospatial data. This software uses conventional methods of pansharpening such as Brovey and IHS to generate a high resolution multispectral image. The time taken to generate such an image ranges from 120 seconds to 300 seconds depending upon the resolution of the panchromatic and multispectral image inputs.

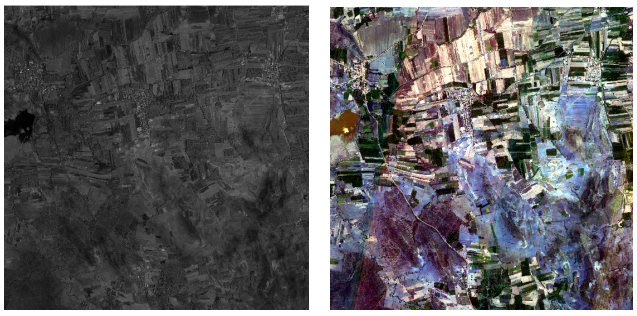


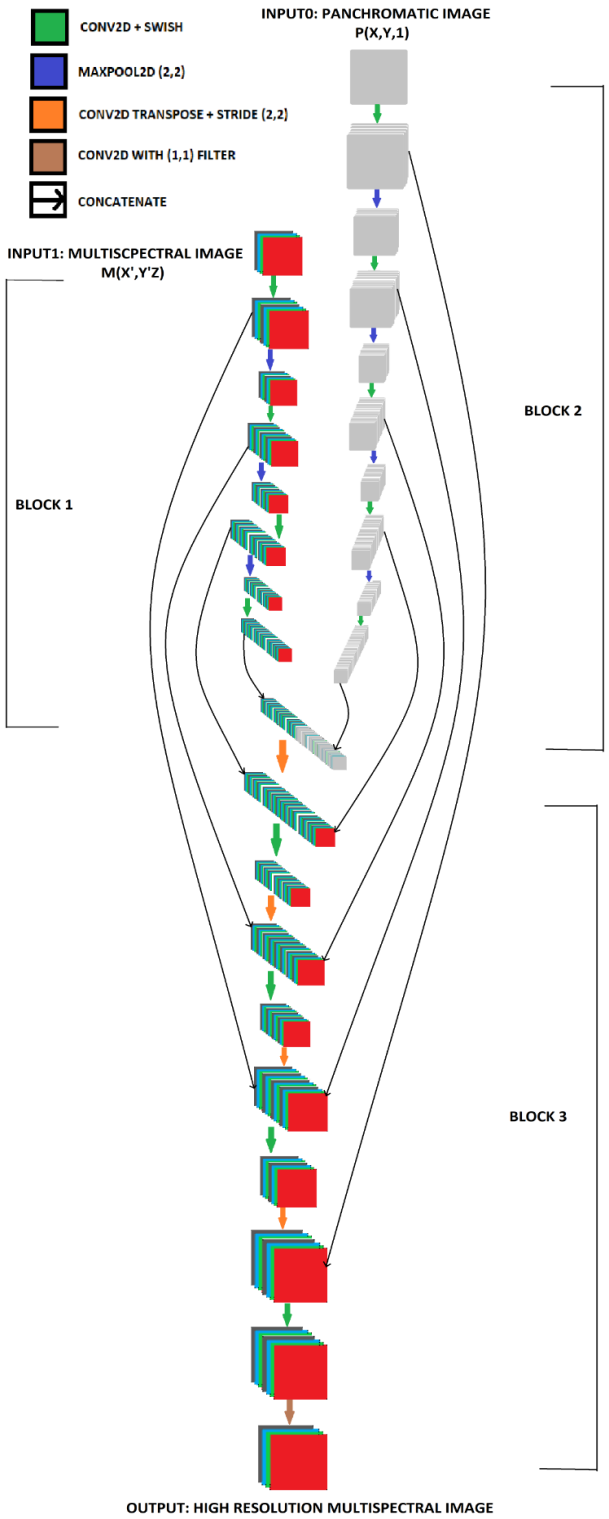
FIG 2: a)PANCHROMATIC IMAGE (9728, 9728) b) MULTISPECTRAL IMAGE (3648, 3648, 4)

2) Training and Results

The original panchromatic, multispectral, pansharepened images are of size (9728, 9728), (3648, 3648, 4) and (9728, 9728, 4) respectively. For training purposes they are broken down into patch images of size (512,512), (192,192,4) and (512,512) respectively. A final predicted image can be obtained by stacking all the predicted images of size (512, 512, 4) in the correct order. The activation function used in the convolutional layers is '*Swish*'. It can be defined as follows:

$$Swish(x) = x * sigmoid(x)$$
$$x \quad : (-inf, +inf)$$
$$Swish \quad : (-inf, +inf) \qquad (eqn\ 1)$$

Swish is a lesser known activation function which was discovered recently, by researchers at Google[5]. It is seen to outperform *ReLU* in most of the Computer Vision related neural networks. The main purpose of using *Swish* is to permit the neurons to have negative weights for normalised input images. *ReLU* has a range of [0, inf) and *tanh* has a range of [-1,1] thus they do not perform well for normalised images having negative pixel values lesser than -1. The range of *Swish* solves the above issue.



OUTPUT(X,Y,Z)

FIG 1: MODEL OF PANSHARPNET

ReLU activation is used in the last output layer. The loss used is Mean Squared Error (MSE) and the optimiser used is Adam

with a learning rate of 0.0005. The data was split into 80% for training and validation, and 20% for testing.
The model was trained with the above mentioned dataset on Google Colab Nvidia K80/T4 12GB GPU. For about 75 epochs, an accuracy of about 0.94 was achieved with training loss of 0.0299. The results of the training can be seen in the TABLE 1, below. There is a drastic improvement in the accuracy within very few epochs. The loss also reduces to a great extent.

Prediction time for the model on the above said hardware is about 0.001 seconds. The output obtained from the model is of the size (512,512,4) which derives its spatial resolution from the panchromatic image input and spectral resolution from the multispectral image input. The output patches are then concatenated horizontally and stacked vertically to build the final output of the size (9728,9728). The final output of the model after stacking can be seen in the FIG 4. This operation takes about 20 seconds to run.

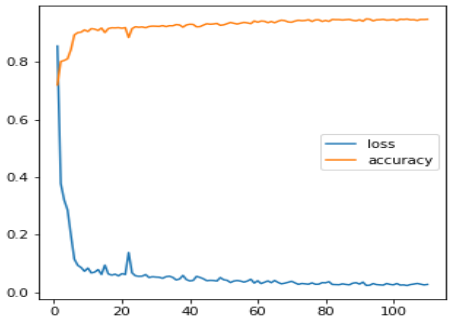


FIG 4: ACCURACY, LOSS VS EPOCHS

The total time cost for the entire fusion process is about 20 seconds. The same procedure cost about 300 seconds on the QGIS software. Therefore, the model successfully achieves its objective by decreasing the time taken for pansharpening drastically, along with a good accuracy.

TABLE 1: TRAINING ACCURACY AND LOSS

EPOCHS	TRAINING RESULTS	
	ACCURACY	*MSE LOSS*
5	0.7955	0.2898
15	0.9213	0.0637
25	0.9220	0.0620
35	0.9193	0.0682
45	0.9238	0.0453
55	0.9354	0.0381
65	0.9386	0.0361
75	0.9424	0.0299

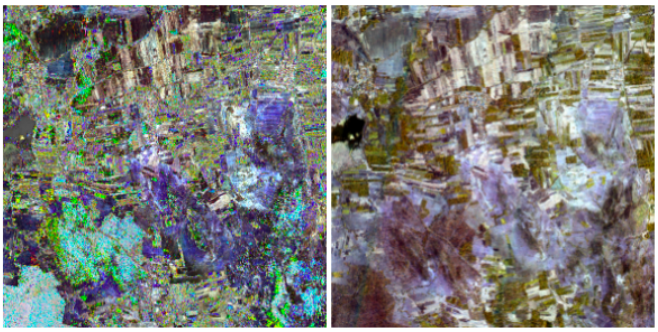


FIG 5: a) MODEL OUTPUT (9728, 9728) b) GROUND-TRUTH QGIS (9728, 9728, 4)

IV. PROPOSED BINARY AND MULTICLASS SEGMENTATION MODEL FOR MULTISPECTRAL IMAGERY

Binary segmentation is performed to detect vegetation in satellite images. Multiclass segmentation is performed to segment the satellite image into 5 classes- vegetation, water, road, urban and water.

1) U-Net

The design of the model includes an encoder and a decoder. It comprises the continuous application of two convolutions of size (2,2). Every layer is trailed by a *ReLU* activation and batch normalization. A max pooling layer of stride (2,2) is applied after the second convolution. After every downsampling operation the total feature channels are increased by doubling them. Each operation in the expansive path comprises an upsampling of the feature maps trailed by a convolution layer of size (2,2) that reduces the total feature channels by half. This is followed by a concatenation operation with the cropped feature map generated, and two convolutions of size (3,3), with a *ReLU* activation and batch normalization. At the last layer a convolution of size (1,1) is used to generate a mapping between every feature vector to the required number of classes. The encoder and decoder paths are symmetric in structure, and hence this results in an architecture that is U-shaped [6].

2) U-Net with a pre-trained encoder

A transfer learning approach is implemented. The ImageNet dataset serves to design a classification model. This learning is used in a semantic segmentation model. A MobileNetV2 model was used for the classification. The pre-trained model is available as part of Keras. The pre-trained MobileNetV2 model acts as an encoder in the revised U-Net model. This part of the model reduces the dimensionality of the image. The expansive path of the model is similar to the original U-Net architecture.

3) ResUNet

The ground truth masks , generated using object based image analysis , are used to train the Deep Residual U-Net (ResUNet) model [13]. The components constituting a

ResUNet model are encoding network , decoding network and a bridge connecting the two networks. The U-Net uses two convolutions of size (2,2), where each is trailed by a *ReLU* activation function. Whereas in ResUNet, these layers are substituted by a pre-activated residual block. The encoder is composed of three encoder blocks. Each of these encoder blocks are constructed using pre-activated residual blocks. The result of each encoder block functions as a skip connection for the corresponding decoder block. The bridge network is also composed of a pre-activated residual block with a stride value of 2. The decoder comprises three decoder blocks, and after each block, the spatial components of the feature map are increased two-folds and the number of feature channels is reduced.

A. Experiments and discussion

1) Dataset Description

a) Bangalore Urban Multispectral Satellite Images dataset: The dataset consists of 415 satellite images and ground truth masks covering the vegetation for the respective ground truths as shown in FIG 5a and 5b. Each is of size 513 x 513 x 3. The ground truth masks are binary masks. It's matrix representation contains a one for pixels which represent vegetation, or a zero if not. When a satellite image is provided as input, the segmentation model outputs a predicted mask. The ground truth images and masks are resized to a size of 256 × 256 × 3 and normalized.

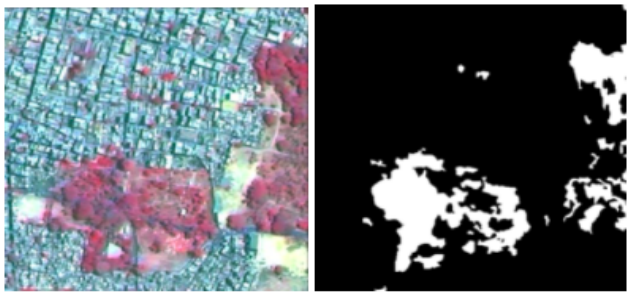


FIG 5: a) SATELLITE IMAGE b) GROUND TRUTH MASK FOR VEGETATION COVER

b) Satellite image of Bangalore: The image in FIG 6a has been processed using object based image analysis to produce its corresponding ground truth masks.

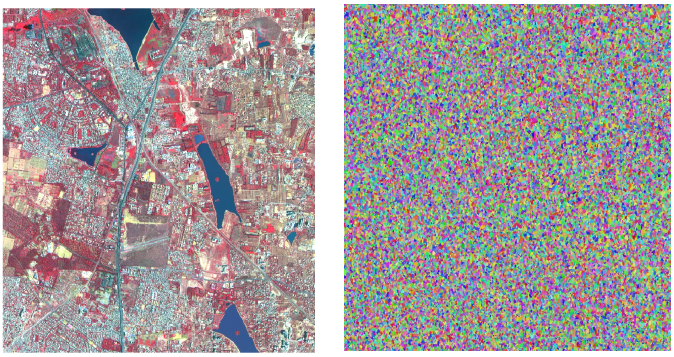


FIG 6: a)SATELLITE IMAGE OF BANGALORE b)SEGMENTED IMAGE OBTAINED USING SLICK

Object based Image analysis (OBIA): It includes pixels firstly being gathered into objects which are dependent on various factors related to spectral similarity of the pixels. A lot of these factors may be estimated, classified as spectral, shape and neighbourhood like mean, standard deviation, size perimeter, compactness, etc. Every object is also part of a 'super-object', acquired by joining a few adjoining objects into one bigger, and each can be partitioned into smaller objects: 'sub-objects'. Utilizing OBIA, information on a land cover image might be incorporated by presenting rules. Whenever a thicket, water and houses are found together, it most probably belongs to an urban area. Similarly when a lot of trees are found, it is likely a forest. This technique achieves accuracies which normal spectral analysis of land cover images cannot. The grouping of pixels into objects can be accomplished using any of the clustering algorithms like slick, quickshift, Felzenszwalb's method or compact watershed method [12].

Slick: This algorithm is primarily inspired by the K-means clustering algorithm. Because of the simplicity of the algorithm, it is very efficient. Slick mainly uses two parameters: compactness is the trade off between color-similarity and proximity, and number of centroids for the kmeans. Here, slick has been used to segment the images into 3,00,000 objects as shown in the FIG 6b as it gives a better representation of the image than quickshift.

The satellite image of Bangalore is partially annotated using QGIS as shown in FIG 7a. Five classes have been considered for segmentation. They are open land, road, urban, water and vegetation.

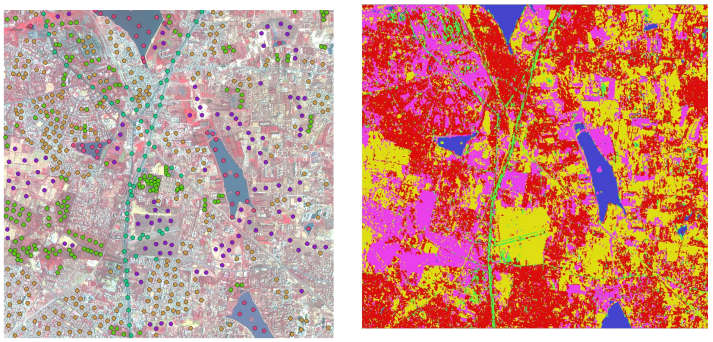


FIG 7: a)MANUALLY ANNOTATED SATELLITE IMAGE OF BANGALORE b)GENERATED GROUND TRUTH MASKS

For each of the segments in FIG 6b, min, max, mean, variance, skewness, and kurtosis are calculated for further analysis. Using these statistics and the manually labelled pixels in FIG 7a as training data, Random Forest classifier is used to classify each of the 300000 objects in FIG 6b into one of the 5 classes. The mask in FIG 7b and the satellite image of Bangalore is broken down into 512 X 512 tiles thus forming a dataset of 120 satellite images.

2) Training and results

The model was trained on the above mentioned datasets on Google Colab Nvidia K80/T4 12GB GPU. A learning rate of 0.0001 was applied. Both the models were trained with a batch size of 32 for binary segmentation and a batch size of 8 for multiclass segmentation. Adam optimizer is used. The

metrics considered is the dice coefficient. The *Loss* is set to *1-Dice*.

TABLE 2: TRAINING RESULTS FOR BINARY SEGMENTATION

MODEL	*EPOCHS*	*PRECISION*	*RECALL*	*DICE-COEFFICIENT*	*LOSS*
U-Net	200	0.8578	0.8449	0.8333	0.1650
	400	0.8723	0.8256	0.8403	0.1585
	600	0.8748	0.8141	0.8378	0.1611
	800	0.8604	0.8507	0.8482	0.1508
Modified U-Net	200	0.8273	0.8462	0.6610	0.3068
	400	0.8083	0.8962	0.6993	0.2704
	600	0.8117	0.9055	0.7166	0.2542
	800	0.8570	0.8795	0.7768	0.2002
	1000	0.8691	0.8842	0.8429	0.1384
	1200	0.8758	0.8797	0.8709	0.1291

After a certain number of epochs, the dice coefficient of the U-Net model did not improve. In the modified U-Net model, there was a steady, but slow increase in the dice coefficient as the number of epochs increased. After the 600th epoch, the batch size was reduced to 8 for both the models and the learning rate was increased to 0.001. This showed no significant improvement in the U-Net model. But the dice coefficient of the modified U-Net model improved at a faster rate as shown in TABLE 2. The U-Net model achieved better results at a less number of epochs. A better overall performance was achieved by the modified U-Net model. But this was possible only with more training. The segmentation results of the U-Net model is shown in FIG 12 and the results of the modified U-Net model is shown in FIG 13.

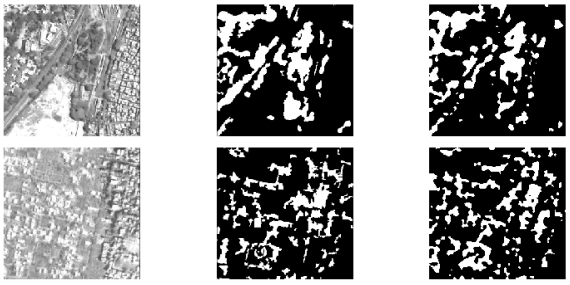


FIG 8: BINARY SEGMENTATION WITH U-NET (satellite image, ground truth mask, predicted mask)

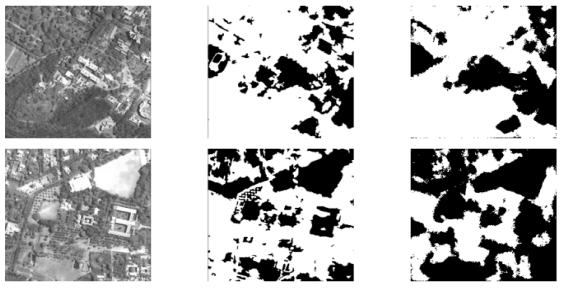


FIG 9: BINARY SEGMENTATION WITH MODIFIED U-NET (satellite image, ground truth mask, predicted mask)

TABLE 3: TRAINING RESULTS FOR MULTICLASS SEGMENTATION

MODEL	*EPOCHS*	*PRECISION*	*RECALL*	*DICE-COEFFICIENT*	*LOSS*
U-Net	50	0.7891	0.6892	0.6464	0.8070
	100	0.8015	0.6369	0.5971	0.7025
	150	0.8062	0.6816	0.6337	0.6725
	200	0.8049	0.6577	0.6035	0.6870
ResUNet	50	0.7424	0.6091	0.5762	0.7712
	100	0.4670	0.4374	0.4202	2.6383
	150	0.8020	0.6933	0.6387	0.6421
	200	0.8118	0.6689	0.6179	0.6912

The ResUNet model performed better than the U-Net model as seen from TABLE 3. The reason why ResUNet succeeds is because it solves the problem of exploding gradients which is a common problem while training satellite images. The skip connections improve connectivity between various layers of the ResUNet, thus facilitating better flow of information between layers. The segmentation results of the U-Net model is shown in FIG 14 and the results of the ResUNet model is shown in FIG 15.

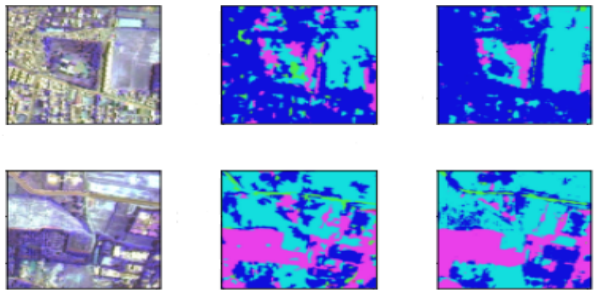


FIG 10: MULTICLASS SEGMENTATION WITH U-NET (satellite image, ground truth mask, predicted mask)

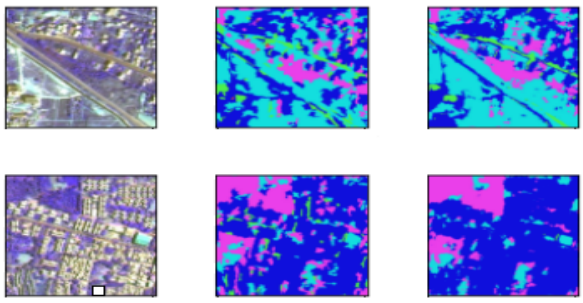


FIG 11: MULTICLASS SEGMENTATION WITH RESUNET (satellite image, ground truth mask, predicted mask)

V. PROPOSED 3D HYPER UNET MODEL

3D convolutions is a relatively new technique used in the domain of hyperspectral data classification, however, there is no dedicated model for segmentation tasks of HSI data. The 3D Hyper-UNET model proposed here is partly inspired from

the UNET architecture, with an additional feature of 3D convolution and the 3D transpose convolution layers. The architecture comprises an input layer with a dimension of H(X,Y,Z) where X*Y represents the spatial resolution and Z represents the spectral resolution. This is followed by two 3D conv layers with filter size of (3,3,3) and a 3D maxpool layer with a pool size of (2,2,2). The output dimension of this layer is H(X/2,Y/2,Z/2). The same combination of 3D conv and 3D maxpool layers repeats three times. Most of the features would have been extracted from the input for classification in the downstream part of the model. These extracted features are then combined and transposed to build back the segmentation masks for the HSI input using 3D conv transpose layers.
Additionally, there are concatenate layers coming from the downstream part of the model, which adds the learnt features to the upstream, thus the learning process becomes easier and quicker. The last layer uses a softmax activation and 1D-2D convolution to give an output of M(X,Y,C) where X*Y represents the spatial resolution and C represents the number of classes involved in the dataset.

$$Model(H(X,Y,Z)) = M(X,Y,C) \quad (eq\ 2)$$

*X * Y -Spatial resolution*
Z - Number of Spectral Bands in HSI
C - Number of Class layers

A. Experiments and discussion

1) Dataset Description

Hyperspectral data for classification and segmentation is obtained from Indian Pines Dataset - 200 bands, 16 classes (TABLE 4), University of Pavia Dataset - 100 bands, 9 classes (TABLE 5) and Salinas Scene Dataset - 224 bands, 16 classes (TABLE 6)

TABLE 4: INDIAN PINES CLASSES

#	*CLASS*	*SAMPLES*
1	ALFALFA	46
2	CORN-NOTIL	1428
3	CORN-MINTIL	830
4	CORN	237
5	GRASS-PASTURE	483
6	HAY	730
7	OATS	28
8	SOYABEAN-NO TILL	478
9	SOYABEAN-MINTILL	20
10	SOYABEAN-CLEAN	972
11	WHEAT	2455
12	GRASS TREES	593
13	WOODS	205
14	BUILDINGS-GRASS	1265
15	STONE-STEEL-TOWERS	386
16	GRASS-PASTURE-MOWED	93

TABLE 5: PAVIA UNIVERSITY CLASSES

#	*CLASS*	*SAMPLES*
1	ASPHALT	6631
2	MEADOWS	18649
3	GRAVEL	2099
4	TREES	3064
5	PAINTED METAL SHEET	1345
6	BARE SOIL	5029
7	BITUMEN	1330
8	SELF-BLOCKING BRICKS	3682
9	SHADOWS	947

TABLE 6: SALINAS SCENE CLASSES

#	*CLASS*	*SAMPLES*
1	BROCOLI_GREEN_WEEDS_1	2009
2	BROCOLI_GREEN_WEEDS_2	3726
3	FALLOW	1976
4	FALLOW_ROUGH_PLOW	1394
5	FALLOW_SMOOTH	2678
6	STUBBLE	3959
7	CELERY	3579
8	GRAPES_UNTRAINED	11271
9	SOIL_VINYARD_DEVELOP	6203
10	CORN_GREEN_WEEDS	3278
11	LETTUCE_ROMAINE_4WK	1068
12	LETTUCE_ROMAINE_5WK	1927
13	LETTUCE_ROMAINE_6WK	916
14	LETTUCE_ROMAINE_7WK	1070
15	VINYARD_UNTRAINED	7268
16	VINYARD_VERTICAL_TRELIS	1807

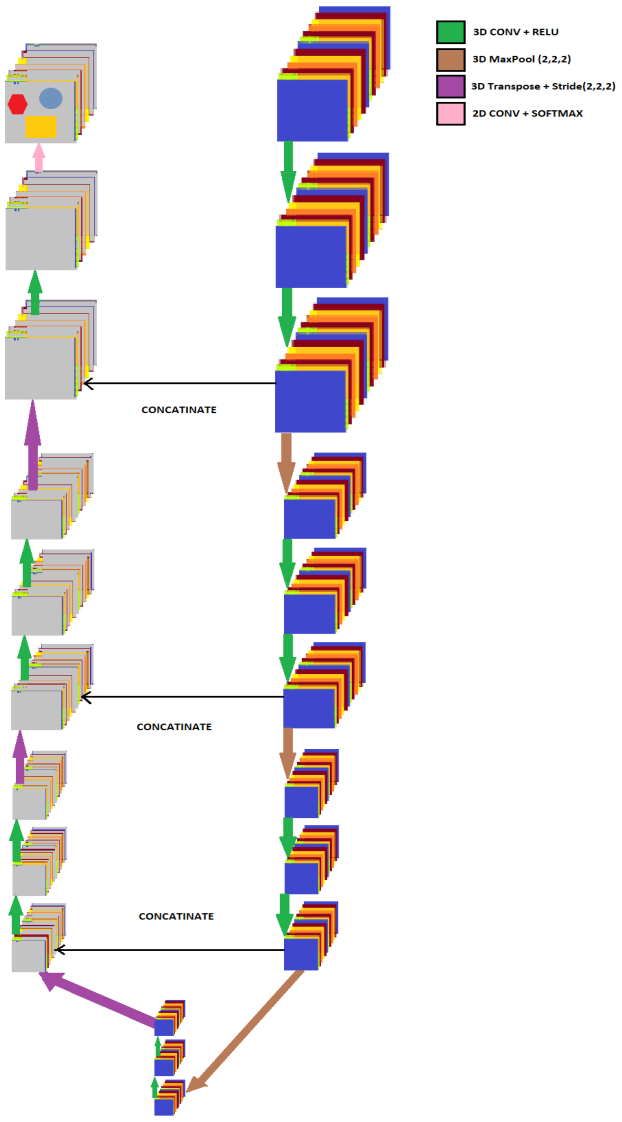


FIG 12: MODEL OF 3D-HYPER UNET

2) Training and results

The activation function used is ReLU for each convolution layer, and softmax layer at the last to build the segmentation masks. The loss used is categorical cross entropy and the optimiser used is Adam with learning rate of 0.001. The model was trained on the above mentioned datasets on Google Colab Nvidia K80/T4 12GB GPU. Intersection over Union(IoU) metric was used for evaluation of segmentation results.

$$IoU = Area\ of\ Overlap\ (AO)\ /\ Area\ of\ Union\ (AU) \quad \text{(eq 3)}$$

TABLE 7: TRAINING RESULTS OF 3D-HYPER UNET MODEL

Dataset	Total Epochs	Training loss	Mean_IOU
Indian Pines	400	0.0698	0.5650
University of Pavia	500	1.1069e-04	0.7204
Salinas Scene	300	2.47e-04	0.7900

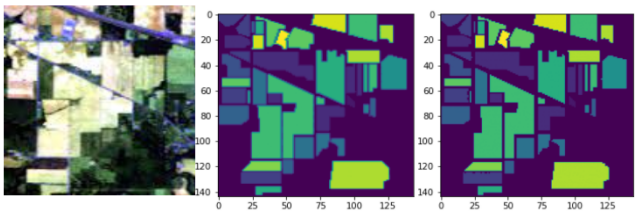


FIG 13: INDIAN PINES HS MULTICLASS SEGMENTATION WITH 3D-HYPER UNET (satellite image, ground truth mask, predicted mask)

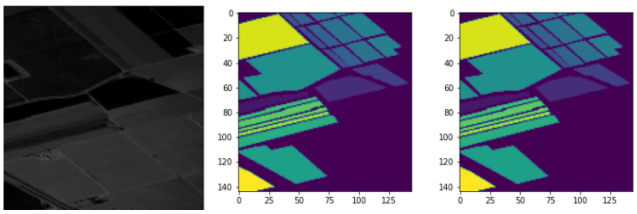


FIG 14: SALINAS SCENE HS MULTICLASS SEGMENTATION WITH 3D-HYPER UNET (satellite image, ground truth mask, predicted mask)

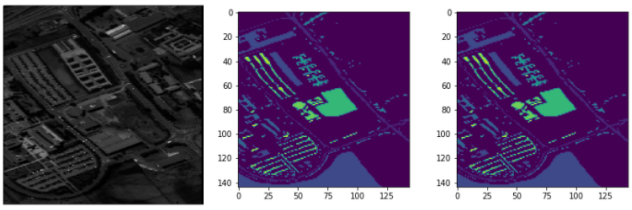


FIG 15: PAVIA UNIVERSITY HS MULTICLASS SEGMENTATION WITH 3D-HYPER UNET (satellite image, ground truth mask, predicted mask)

VI. CONCLUSION

The pansharpening CNN architecture - PansharpNet proposed in this paper explores the usage of convolutions and parent features concatenation with appropriate activation functions for pansharpening of multispectral images. The model utilizes all the bands of multispectral image and gives a high resolution output with a good accuracy. It is seen to be computationally efficient with low MSE loss on the datasets tested. Thus the proposed model is novel and can be considered as the state of the art on pansharpening of multispectral images. In the future research will be carried out on a similar model for hyperspectral images with 3D CNN as the features of all the bands are supposed to be considered.

A U-Net architecture and its variations have been proposed for binary semantic segmentation for detection of vegetation and multiclass semantic segmentation for detection of classes like vegetation, water, land cover, road and urban area using few data. A comparative result was provided between two different models for each type of segmentation. It was seen that the variations of the U-Net model performed better than the original architecture.

The 3D Hyper-UNET architecture proposed in this paper explores the possible usage of 3D convolutions for hyperspectral image classification and segmentation. The model explores the spatial and the spectral information from the HSI using 3D convolutions to perform better and accurate segmentations. The model performs a one step segmentation process, unlike the predecessors who mainly concentrate on classification followed by pixel wise colouring via an iterative method for generation of masks.

REFERENCES

[1] Parente C.*, Pepe M.Department of Sciences and Technologies, University of Naples –Italy,Influence of the weights in IHS and Brovey methods for pan-sharpening,International Journal of Engineering & Technology,July 2017.

[2] Jaewan Choi , Honglyun Park and Doochun Seo, Chungbuk National University, and Korea Aerospace Research Institute,Pansharpening Using Guided Filtering to Improve the Spatial Clarity of VHR Satellite Imagery,*Remote Sens.* 2019,15 March 2019.

[3] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva and Giuseppe ScarpaDIETI, University Federico II of Naples, Italy,CNN-BASED PANSHARPENING OF MULTI-RESOLUTION REMOTE-SENSING IMAGES,*Remote Sens.* 2016,14 July 2016.

[4] Zhiguang Yang, Youping Chen, Zhuliang Le & Yong Ma ,GANFuse: a novel multi-exposure image fusion method based on generative adversarial networks,*Neural Computing and Applications* 2020.

[5] Prajit Ramachandran* , Barret Zoph, Quoc V. Le , google brain,Searching for activation functions,october,2017

[6] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", Computer Science Department and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany, 2015.

[7] Guillaume Chhor, Cristian Bartolome Aramburu, Ianis Bougdal-Lambert, " Satellite Image Segmentation for Building Detection using U-net", 2017.

[8] Ulmas, Priit and I. Liiv, "Segmentation of Satellite Imagery using U-Net Models for Land Cover Classification." ArXiv abs/2003.02899, 2020.

[9] Alhassan, Victor, "A deep learning framework for land-use/land-cover mapping and analysis using multispectral satellite imagery." Neural Computing and Applications 32, 2019: 8529-8544.

[10] Qishuo Gao , Samsung Lim and Xiuping Jia, "Hyperspectral Image Classification Using Convolutional Neural Networks and Multiple Feature Learning"

[11] Swalpa Kumar Roy, Gopal Krishna, Shiv Ram Dubey, Member, IEEE,and Bidyut B.Chaudhuri, Life Fellow, IEEE"HybridSN: Exploring 3D-2D CNN FeatureHierarchy for Hyperspectral Image Classification".

[12] T. Blaschke, Object based image analysis for remote sensing, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 65, Issue 1, 2010.

[13] Diakogiannis,"ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data." ArXiv abs/1904.00592, 2019.