BUILDING FOOTPRINT EXTRACTION FROM SPACE- BORNE IMAGERY USING DEEP NEURAL NETWORKS

Banda Tejeswari [1], Surendra Kumar Sharma [1*], Minakshi Kumar [2], Kshama Gupta [1]

[1] URSD, Indian Institute of Remote Sensing, Dehradun, Uttarakhand, India- (bandatejaswari@gmail.com, ssharma3@ce.iitr.ac.in, gupta.kshama@gmail.com)

[2] PRSD, Indian Institute of Remote Sensing, Dehradun, Uttarakhand, India - minakshikumar900@gmail.com

Commission II, WG II/6

KEY WORDS: Deep Learning, Building foot-pint extraction, Mask RCNN, Open Source training data, Open Street Map, Remote Sensing, Satellite imagery.

ABSTRACT:

One of the important and high-level detailing contained within basemaps is the ‘building feature’. Though pre-trained Deep Learning (DL) models are available for Building Feature Extraction (BFE), they are not efficient in predicting the buildings in other locations. This study explores the need and the major issue of implementing DL models for BFE from Very High Resolution Remote Sensing (VHRS) satellite data for any given area. Though advanced DL models are invented, in order to implement them, huge amount of potential training data is demanded for feed in. the building typologies are highly subjected to the context of study area including soil characteristics, culture/lifestyle/economy, architectural style and the building byelaws. The study believes that availability of enough training data of contextual buildings as one of the concern for effective model training. The study aims to extract the buildings present in the study area from Pleiades 1A (2019) RGB VHRS data using simple Mask R-CNN instance segmentation model which is training on the native contextual buildings. Here, an automated method of generating the location-specific training data for a given area is followed using Google Maps API (2021). The generated training data when trained on a deep learning architecture and predicted by the input data yielded promising results. The prediction accuracy of about 98.41% specificity, 96.20% predictive accuracy and 0.89 F1 score are achieved. The methods adopted assist the planning/governing bodies to accelerate the qualitative urban map preparation.

1. INTRODUCTION

Buildings are basic landscape features that form the urban fabric. Capturing and including buildings in map making process contribute to rich database that supports inter-disciplinary and micro-urban studies. Extracting building features from optical remote sensing images has always been an active research in the field of computer vision.

Techniques of feature extraction from satellite imagery mainly depend on the characteristics of feature, scene complexity, dataset quality (complexity and scale), context of the target area and the application of output. There are four broad methods of feature extraction from imagery dataset. They are pixel based classification, traditional object detection, the machine learning and the deep learning techniques. Pixel based method is mainly used to classify the built-up land cover from medium to coarse resolution images. This approach is followed when the feature size is negligible than that of the pixel size. Object detection method that deals with segmented objects and classifies them using feature properties in the image such as textural, contextual, spectral and geometry. This method is followed when the feature size is larger than that of the pixel size. Machine learning is a semi-automated and supervised learning technique that uses feature descriptors for recognition and classification algorithms to extract the features form the image. Deep learning is a recent, automated and data aided technique that uses heavy convolution layers to extract the features in a single step.

A novel study by (Wang et al., 2014) performed a unique structuring element and had considered the statistical standard deviation of pixels within the kernel for differentiating the regions and the ridges within the multi-scale wavelet images, in order to have an efficient colour and texture-based multi-scale image segmentation.

Researchers have also applied morphological Top-hat filters and K-means algorithm for automatic extraction of building footprints from VHRS data (Gavankar et al., 2018). The uniqueness of this study is that it has firstly thresholded the image into three parts (dark, medium and bright), and then extracted the buildings present within each threshold range so that the buildings with lesser spectral response can be identified and extracted separately. It further uses the dimensional properties to avoid the falsification of buildings.

Building features don’t have the same spectral property always as they vary due to the age, the contrasting background, the colour of the roof, size, texture, etc. (Figure 1) Building features of the same location and from same source show variation as per built properties and the spatial arrangement. However, it is

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-641-2022 | © Author(s) 2022. CC BY 4.0 License.

found that building features, though have varied shape, size and texture, they possess similar contrast and brightness as compared to their adjacent features (Gavankar et al., 2018) and hence, top-hat morphological transform serves well for BFE. However, this algorithm found to be more effective in low-building dense areas that are surrounded by the vegetation.

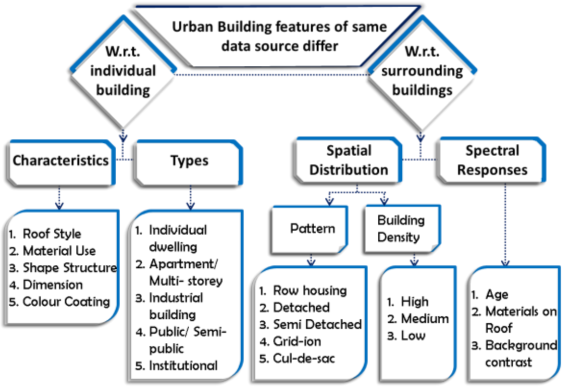


Figure 1. Factors that cause variation in building features within the same location

Building feature form spatial data meant the footprint of the building rooftop. The buildings possess variety of rooftops as per physical built characteristics of the building and its usage. Since the task of feature extraction also depends on the ground complexity, arrangement of the building features and the surrounding landscape features must also be considered. Hence the housing pattern, built density and contrast come into play.
In order to conduct micro level urban analysis (such as thermal demand versus the layout planning & height regulations), it becomes essential to extract the individual building footprints within the given area. The studies done on BFE using space-borne VHRS images adopting OBIA and machine learning techniques have the following limitations.

- Each feature descriptor recognizes a specified texture/geometry/scale of the feature. BFE using single machine learning/object based method doesn't extract all types of textured and colored buildings.
- Buildings are minute features in VHRS dataset. Extracting individual building footprints using the above techniques is feasible only in the cases of low density with distinguishable contrasting surrounding features. It becomes extremely difficult to obtain a definite building boundary in the cases of high density areas with similar contrasting surrounding features. In fact, in urban context, the roads abutting to buildings show similar spectral response. Hence in most cases, the building features are extracted in groups instead being distinct.
- Also, the roof design and the materials present on the roof contribute to the building boundary leakage which makes it hard to obtain the definite building boundary.

Deep learning technique for BFE which has been an active application since three years is found to overcome the above limitations and yield satisfactory results.

DL based semantic segmentation is proposed for BFE from VHRS in study (Li et al., 2019). In this study, huge training data of Worldview-3 from SpaceNet (VanEtten et al., 2018), along with the openly available building information from Google Maps, OpenStreetMap and MapWorld are fused together and augmented while post processing for boosting up the output accuracy. The study uses U-Net model which was performed on four cities i.e, Las Vegas, Paris, Shanghai and Khartoum and obtained precise building outlines with higher F1-scores 0.8911, 0.7555, 0.6266, and 0.5415, respectively.

The study (Roscher et al., 2020) considered the input raster (Worldview II, 1PAN and 8MS bands) with 10,000 precisely digitized and the annotated building labels that covered 25% of the same study area as the input training data (and testing data) to the deep learning model and performed the instance segmentation (Mask-RCNN model) to predict and extract the remaining 75% of the building features within the study area. The training data contained mix of industrial and residential area with all land-use buildings and was considered as one of the efficient benchmark data (for ground truth training and testing of other algorithms) for old town of Semcity, Toulouse.

In DesnseNet architecture, the feature reuse is made between each feature map in iterative concatenation. The study by (Yang et al., 2018) applies and evaluates four DesnseNet models i.e, Branch-out CNN, FCN, CRFasRNN and SegNet, on aerial images of 1M for multiscale semantic segmentation. The training data consists of 4000 RGB&NIR images selected from various locations across the country. In order to further improve the BFE at instance level and to identify the state-of-art CNN, the study proposed 9 different models using the above combinations of pre-trained CNN models, binary labelling, distance-transformed labelling and NIR band fusion. FCN-4s-Bin, FCN-8s-Bin, FCN-4s-CRF-Bin, FCN-8s-CRF-Bin, SegNet-Bin, SegNet-Dist, SegNet-Bin-Fused, SegNet-Dist-Fused and 3Conv-Dist are building extraction models are proposed, predicted on huge scale (entire US) and tested on 78 spatially distributed sites for accuracy comparison. High Performance Computing (HPC) systems with multi-GPU memory was utilized for faster generation of large scale and accurate building maps.

Building footprints in Yangon City are extracted by conditional GAN which is image to image transformation technique and is trained on image pairs (Aung et al., 2020). The input training and testing datasets are taken from GeoEYE (monocular optical RGB images) of Dagon Township. By changing the model parameters (learning rate, b1 –Adam and number of filters in initial convolution layer of generator and discriminator model), 8 different training models were developed. The obtained output images of these models are converted to vector format for accuracy estimation with manually digitized polygons. The results of BFE show 71% of completeness, 81% of correctness and 69% of F1 score. The insights made from the study are 1. If the spectral properties of the buildings in the validating/predicting datasets are similar irrespective of the city, this cGAN pix2pix model, with same training parameters yields better feature extraction. 2.Though the rooftops are diverse (in terms of color and geometry), training pix2pix model with images from the given study area (i.e., features of same predictable area) would produce promising results.

Urban cities are often congested in its downtown area. The building features are tightly packed with haphazard arrangement which makes BFE difficult. This bottleneck was attempted by using improved Mask R-CNN framework which detects the distinct building boundaries even in complex target area (Wen et al., 2019). The refined Mask R-CNN initially fixes the bounding boxes of the detected buildings as per the minimum enclosing rectangles. Later, through rotation anchor (in RPN), these boxes are inclined along their principal directions (w.r.t to aspect ratio). Then in ROI align, after the anticlockwise rotation, feature regions are said to pass through multi-branch prediction network where additional RFB modules (using atrous convolution and inception block) are incorporated to the segmentation branch to deal multi-scale variability. In the end, the obtained rectangle bounding boxes are rotated clockwise. Large set of openly available Google Earth images is used as training data.

Inferring from the literature, the major insights for effective deep learning based feature extraction are, 1. These models are driven by large amount of training data. 2. The building typologies are highly subjected to the context of study area including soil characteristics, culture/lifestyle, architectural style and the building byelaws. The main reason that DL technique yields significant results (in case of building extraction) is that the trained model not only learns to detect the building, but also learns the consistency/ diversity of building features within a region, the similarity between buildings and the background. (This is also a reason why the direct application of pre-trained models doesn’t serve the purpose). Since the training retains the contextual properties of the feature, it is important to ensure that training, validating and prediction datasets must not have large variation among them.

Lack of enough training data is the reason why most of the cities don’t prefer application of DL technique. Creating a large amount of appropriate training data is really challenging task. Few studies took the advantage of publicly available data stores (such as SpaceNet, Inria, Kaggle). But these sources only possess datasets of few cities. The model trained on this data is not feasible to extract building in every location due to change in the context and characteristics.

This study aims at generating a deep learning based building footprint extraction model which is trained on the contextual location specific training data using Google Maps. The output model generated is expected to extract the distinct building footprints in Pleiades 1A VHRS RGB data. The current study believes that training with native buildings is found to give better output. Hence the study follows a different approach of generating the location-specific training data of buildings from freely available data source using Google Maps API.

Google maps are found to create potential training database for urban areas. Google Map products contain Google satellite (Raster RGB) (Figure 2) and the building footprint outline in Default Google Map (Figure 3).

Figure 2. Google satellite Basemap layer

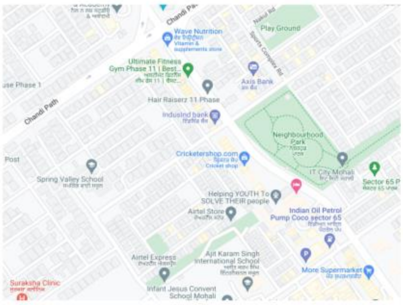


Figure 3. Google Maps Basemap Layer

2. METHODS AND MATERIALS

2.1 Study Area and datasets

A small area of about 26.33 hectare in Mohali region near to Chandigarh is considered as the target area (Figure 4) where the buildings within this site are predicted using the model developed in here.

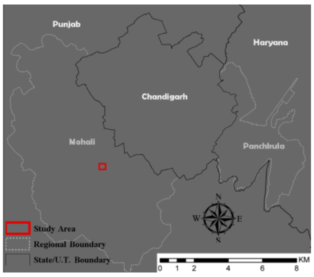


Figure 4. Key map of Study Area

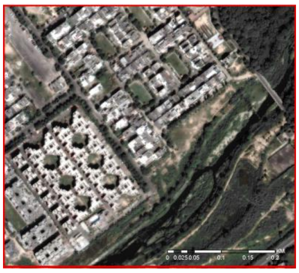


Figure 5. Pleiades pan-sharpened data for target site

Pan-sharpened Pleiades-1A RGB data for this selected site is used for predicting the building footprints (Figure 5). This data is acquired on 10th October 2012 and the spatial resolution of PAN is 0.5M and the 3band RGB MS data is 2M.

2.2 Methodology

Five major steps are involved for creating location-specific training data, BFE model and extraction of building footprints (Figure 6). Initially, sites within the same city region area selected. Then, generation of RGB image and vector label data from Google Maps API in the above sites have been carried out. Further, post processing of the above data for training is performed. Thereafter, the training of the model is done by Mask R-CNN architecture and the above training data. Finally, prediction of this model on Pleiades 1A data for the study area is done followed by accuracy assessment.

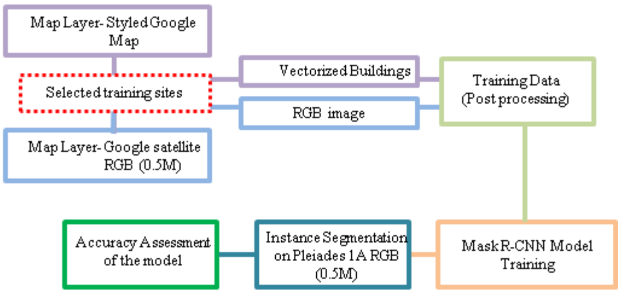

Map Layer- Styled Google Map
Selected training sites
Map Layer- Google satellite RGB (0.5M)
Vectorized Buildings
RGB image
Training Data (Post processing)
Accuracy Assessment of the model
Instance Segmentation on Pleiades 1A RGB (0.5M)
MaskR-CNN Model Training

Figure 6. Methodology

In the very initial step, a manual selection of sites of high building density area with varied building typologies within the same city region is done by keen observation of building properties. These site areas are marked in neat vector polygon layer. Since Chandigarh is the major urban area near to the study area, Chandigarh's sites of higher density, the sites with matching building typologies w.r.t the study area buildings and the sites nearby the study area are chosen and marked.

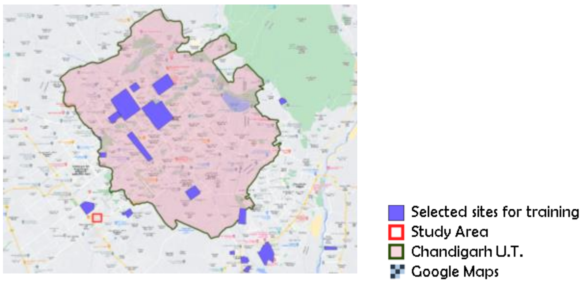


Figure 7. Location of sites selected for creating training data

As Google Maps and Google Satellite are found to be the potential source to obtain the desired training data in the form of matched vector and raster for a given area, the next task is to generate an RGB raster and its corresponding building labels (in vector form) for these selected sites.

The Google Satellite does not permit downloading data from it. However, Google maps API allow its users to create a map layer from it. Using the Google Satellite as the base layer, and the selected sites polygon layer as the mask layer, a map layer is generated for above selected sites for 19- January- 2021. It carries 0.5M 3band RGB information of the selected sites (Figure 8). The next step is to get the vector layer form Google Maps using its API. Although the vector tile is not allowed for downloading, the recently updated Google maps API allows its user to style the map and create a map layer.

Google Maps (Figure 9) contains huge locational data (such as schools, malls, open areas, water bodies, road network, traffic information, etc.). Styling allows users to simplify and enhance the required landscape features as per the study/application while dealing with the subset of huge content from raster tiles. Here, the styling of Google Maps layer is done to enhance building features present in it. The styled Google map with highlighted building features is now kept as a base layer, and the selected sites polygon layer as the mask layer, a map layer is generated from it (Figure 10). This carries 3band image of styled map with building polylines information for the marked sites. In order to obtain the vector building polylines, this 3band raster is initially said to undergo raster class segregation (background and buildings). Now using ArcScan (raster clean-up and vectorize tools), vectorization of the all the foreground features is performed. Finally, the above vector is converted into polygons and an attribute field of 'building' (annotation) is added. The resultant would be the vector building layer file with annotations (Figure 11- has the annotated vector buildings layer overlaid on Google satellite image).

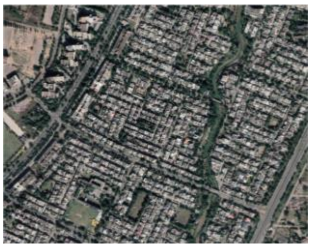


Figure 8. Google Satellite base layer -in High density area

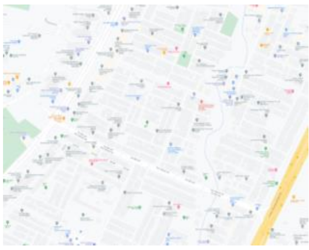


Figure 9. Default Google Map

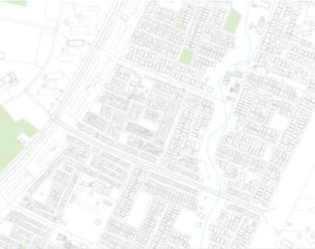


Figure 10. Styled Google Map

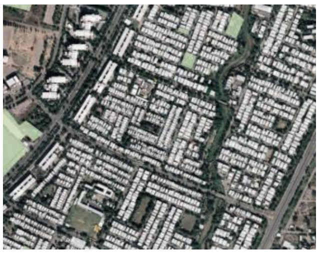


Figure 11. Vectorized building labels and raster map layer

Once, both the raster and the vector building annotation layer is made available, post processing of this data was carried out by proper square tiling.

Below are the characteristics of the images used:

- Total number of bands- 3 (R,G,B)
- Total no. of image grids for site locations – 16,244
- (with augmentation & padding)
- CRS- GCS WGS1984 (epsg-4326)
- Vector Annotation format: Shapefile
- Each image tile – 256 x 256
- Stride – 128 x 128
- Spatial Resolution - 0.5M
- Total no. of feature labels 2,67,342

2.3 Model training

As aimed for obtaining distinct building boundaries, the study uses deep learning based instance segmentation model, Mask RCNN architecture for training. In instance segmentation, each object within the same class will be assigned with an instance. As Mask RCNN, a state-of-art framework which is built on top of Faster R-CNN is one of the recent deep learning models that best support instance object detection (and segmentation), it is chosen for BFE. Mask RCNN = (Faster RCNN) + (FCN Mask Head). FasterRCNN is widely used for object detection. It returns the class label & the bounding box for each individual object that is detected within the image by incorporating an attention mechanism using Region Proposal Network (RPN). Mask RCNN, and additionally it possesses a branch which is responsible for generation an object mask (/segmentation mask) within the detected object region called semantic segmentation (Figure 12). The architecture of any single DL model contains a ResNet50 CNN backbone (Encoder) for feature extraction (based on huge convolution operations) and two Decoders (1FCN classifier-for predicting the classes & Bounding Boxes and 1Mask head for segmenting the objects detected).

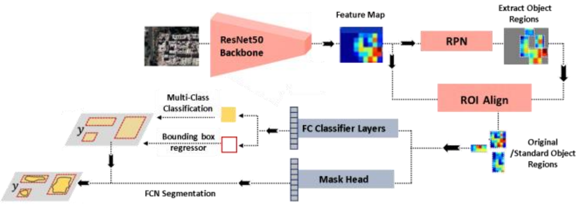


Figure 12. Architecture of Mask-RCNN

In the first stage, FasterRCNN uses ResNet50 (Residual Network with fifty hidden layers in architecture) as one entire CNN backbone model to extract feature map (creating/highlighting the features) from the input image. The convolution blocks in resnet50 architecture run on the logic of switch connections. In Second stage, the task is to determine and create the set of regions/RoI (Region of Interest)/bounding boxes (of the objects) within the feature map using RPN protocol. All the other portions of the image leaving these regions are considered as background and are not processed further in 3rd stage. Faster RCNN, in the context of Mask RCNN, in the third stage, considers both the feature map and the determined regions (foreground features of the above stage) to perform standardization of these regions using ROI Align for better accuracy by preserving the spatial orientation of features with no data loss. In fourth stage, using dense layers on the output from ROI Align, classification is performed to return the class label of the object in each ROI. In the last stage, using FCN mask head, the pixel-based segmentation/object masking is performed on these standardized and classified object regions of the above stage.

3. RESULTS

3.1 Model training

As discussed earlier, Google map layers from Google API has been utilised in this study for training the model (Figure 13). A large amount of training data has been ingested and learning rate of the model is improved with continuous ingestion of training data (Figure 14). It can be seen that with each epoch, training loss and valid loss values kept on improving (Figure 15). Figure 16 shows the finally obtained model results. Left side image shows ground truth data while right side image shows model outputs. It can be seen that there is high correspondence of model outputs with ground truth data.

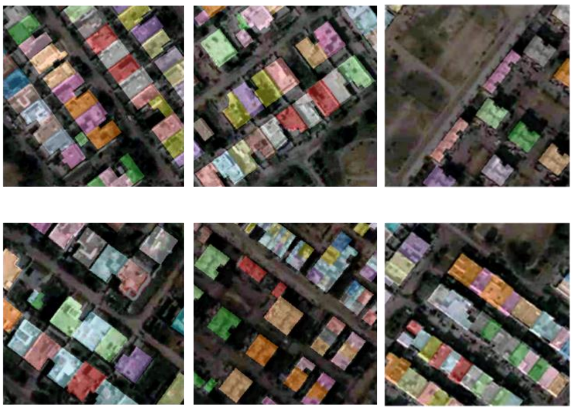


Figure 13. Training data image tiles along with building mask

3.1.1 Trained Model Statistics:

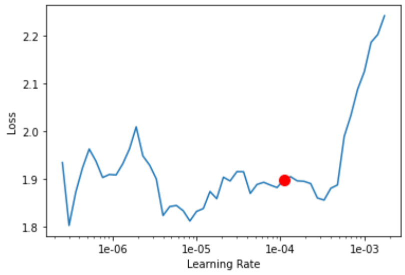
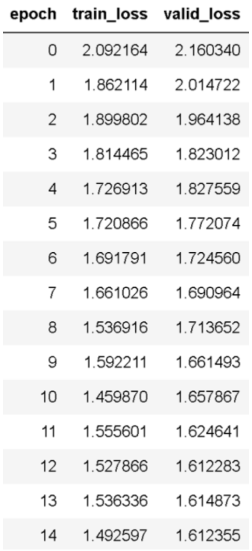


Figure 14. Learning rate of the model



epoch	train_loss	valid_loss
0	2.092164	2.160340
1	1.862114	2.014722
2	1.899802	1.964138
3	1.814465	1.823012
4	1.726913	1.827559
5	1.720866	1.772074
6	1.691791	1.724560
7	1.661026	1.690964
8	1.536916	1.713652
9	1.592211	1.661493
10	1.459870	1.657867
11	1.555601	1.624641
12	1.527866	1.612283
13	1.536336	1.614873
14	1.492597	1.612355

Figure 15. Details of the trained model (epochs, training loss and valid loss)

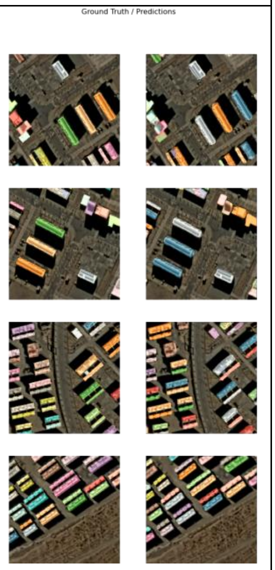


Figure 16. Model results, left sided images represent ground truth data, right sided images represent predicted building footprints

3.2 Building Footprint Extraction

Implementation of model on the entire selected study area results (Figure 17) shows individually extracted buildings with great correspondence with base image. It proves that use of open data sources from OSM (for vector label) and Google earth (as raster) for training can benefit the faster preparation of the efficient training data. The OSM building vector file is available for major cities. Selecting the city nearby the study area and extracting them for training would generate a better training data. This method can be beneficial in creating faster and accurate urban database.

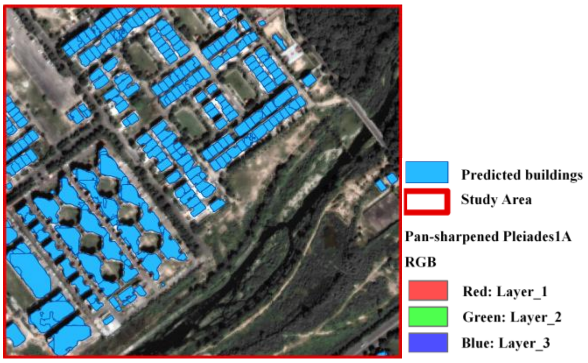


Figure 17. Building footprint extraction on Pleiades 1A RGB data using the Model

3.3 Accuracy Assessment

Accuracy assessment of extracted building footprints with digitized footprints yielded high level of accuracy with 98.41 specificity (Figure 18 and 19). The predictive accuracy is found to be 96.20 with FPR of only 1.58 (Table 1).

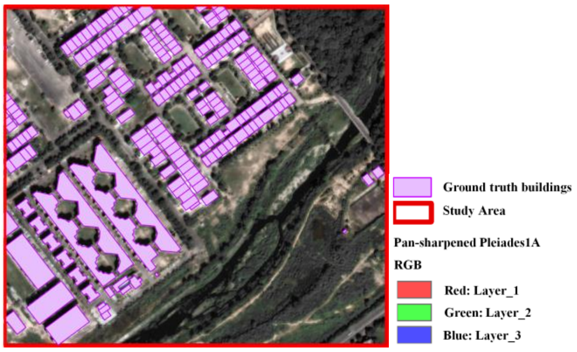


Figure 18. Ground truth buildings

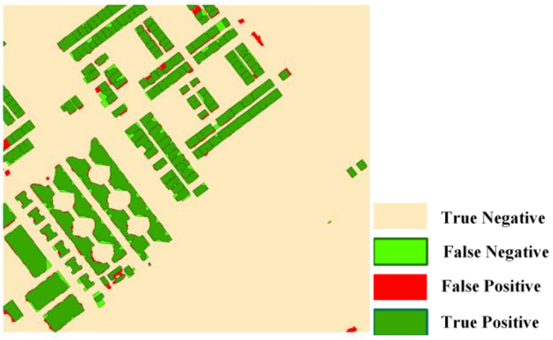


Figure 19. Accuracy estimation

Accuracy Measure	Value
Sensitivity/True Positive Rate (TPR)	86.77
Specificity/TNR	98.41
Predictive Accuracy	96.20
PPV/ Precision	92.77
NPV	96.94
F1-Score	0.89
FPR	1.58
PPR	30.37
NPR	7.45
False Positive Fraction (FPF)	7.22
False Negative Fraction (FNF)	3.05

Table 1: Accuracy estimation

4. CONCLUSION

The study explores different reasons why traditional object based and the machine learning based feature extraction techniques, while dealing with VHRS imageries, have limitation especially in case of buildings. In order to produce high quality urban maps for micro analysis or detailed damage estimation during disaster, 'building footprint' turn to be an inevitable landscape urban features those to be included in the maps. Hence deep learning techniques are found to be a ray of hope to extract definite building features from VHRS satellite data. Though pre-trained DL models are available, they are not efficient in predicting the buildings in other locations. Since being data driven learning technique, one of the greatest bottlenecks for implementing these deep learning models directly on any given area to predict buildings is the training data as there is high subjectivity due to context of study area that includes soil characteristics, culture/lifestyle/economy, architectural style and the building byelaws.

Creating a large amount of appropriate training data is really challenging task as DL methods being data driven learning. This study overcomes this challenge by generating large amount of training data that is trained on the native buildings of the same city region as of the study area. It performs building footprint extraction on 2012 Pleiades 1A data RGB data and entire training data is produced using Google Maps API. Google maps carry building locations for most of the cities worldwide. As Google Maps and Google Satellite are found to be the potential source to obtain the desired training data of matched building polygons and raster for a given area, this advantage is further utilized for effective model generation to detect buildings of a specified place.

The extraction results are promising and the estimates of the accuracy are 98.41% Specificity, 96.20% Predictive accuracy and 0.89 F1 score. In this study, contextual building information is used as training data. Even by using a simple Mask R-CNN technique high accuracies are obtained. Here, 2021 RGB training data is used to predict 2012 RGB buildings. Any other RGB data with similar building properties as that of training data, when predicted by the model, irrespective of timelines and data source, better results are acquired. This can be a greater help in emergency disaster periods to generate instance maps using other VHSR RGB data. The model displayed the high accuracy obtained for this study even in high density built up area of Chandigarh city. It should be noted that

India has high density of built up in some parts of its cities being the second most populous country and large number of urban population. Automatic extraction of building footprints especially in high density built-up is the need of the hour for sustainable and long term planning of these cities. Developing an efficient model for automatic extraction of all the buildings within the given urban area would greatly serve over crowded cities/towns

REFERENCES

Aung, H.T., Pha, S.H., Takeuchi, W., 2020. Building footprint extraction in Yangon city from monocular optical satellite image using deep learning. Geocarto Int. https://doi.org/10.1080/10106049.2020.1740949

Gavankar, N.L., Ghosh, S.K., 2018. Automatic building footprint extraction from high-resolution satellite image using mathematical morphology. Eur. J. Remote Sens. 51, 182–193. https://doi.org/10.1080/22797254.2017.1416676

Li, W., He, C., Fang, J., Zheng, J., Fu, H., Yu, L., 2019. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. Remote Sens. 11, 1–19. https://doi.org/10.3390/rs11040403

Roscher, R., Volpi, M., Mallet, C., Drees, L., Wegner, J.D., 2020. SEMCITY TOULOUSE: A BENCHMARK for BUILDING INSTANCE SEGMENTATION in SATELLITE IMAGES. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 5, 109–116. https://doi.org/10.5194/isprs-annals-V-5-2020-109-2020

VanEtten, A., Lindenbaum, D., Bacastow, T.M., 2018. SpaceNet: A Remote Sensing Dataset and Challenge Series.

Wang, C., Shi, A.Y., Wang, X., Wu, F.M., Huang, F.C., Xu, L.Z., 2014. A novel multi-scale segmentation algorithm for high resolution remote sensing images based on wavelet transform and improved JSEG algorithm. Optik (Stuttg). 125, 5588–5595. https://doi.org/10.1016/j.ijleo.2014.07.002

Wen, Q., Jiang, K., Wang, W., Liu, Q., Guo, Q., Li, L., Wang, P., 2019. Automatic building extraction from google earth images under complex backgrounds based on deep instance segmentation network. Sensors (Switzerland) 19. https://doi.org/10.3390/s19020333

Yang, H.L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., Bhaduri, B., 2018. Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 2600–2614. https://doi.org/10.1109/JSTARS.2018.2835377