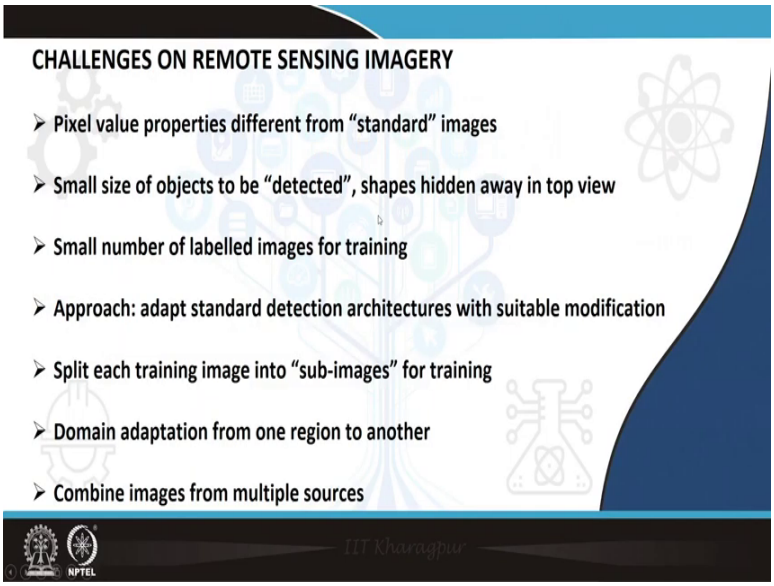
Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

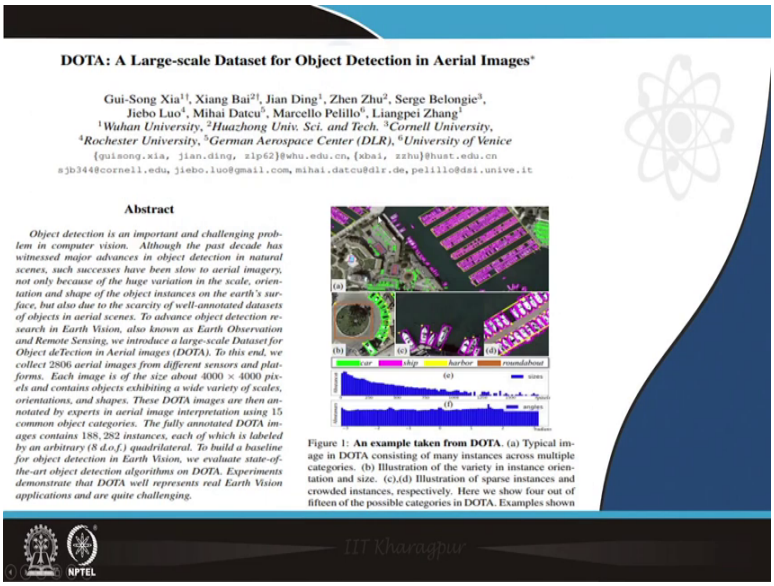
Module - 04
Machine Learning for Earth Observation Systems
Lecture - 28
Object Detection in Satellite Imagery - 2

Hello everyone. Welcome to lecture 28 of this course on Machine Learning for Earth System Science. So, like we are currently in module 4, where we are focusing on Machine Learning for Earth Observation Systems. And, in this lecture we are going to talk about Object Detection in Satellite and in remote sensing Imagery.

(Refer Slide Time: 00:45)



(Refer Slide Time: 00:45)



So, like we had already discussed in the last lecture about how object detection is done using like in standard computer vision using deep learning architectures like R-CNN and YOLO. And, we have also discussed the different challenges of applying those approaches in remote sensing imagery because of the various natures of like remote sensing.

So, now what we are going to do today is or in today's lecture is let us focus on a few research papers where or which focus on detecting different types of objects from different types of remote sensing images. So, here is the first paper which actually builds a data set of these kinds of objects captured from aerial images. This data set is called DOTA.

So, object detection in an is an important and challenging problem in computer vision. Although, the past decade has witnessed major advances in object detection in natural scenes, such successes have been slow to aerial imagery, not only because of the huge variation in the scale, orientation and shape of the object instances in the on the earth surface, but also due to scarcity of well annotated data sets of objects in aerial scenes.

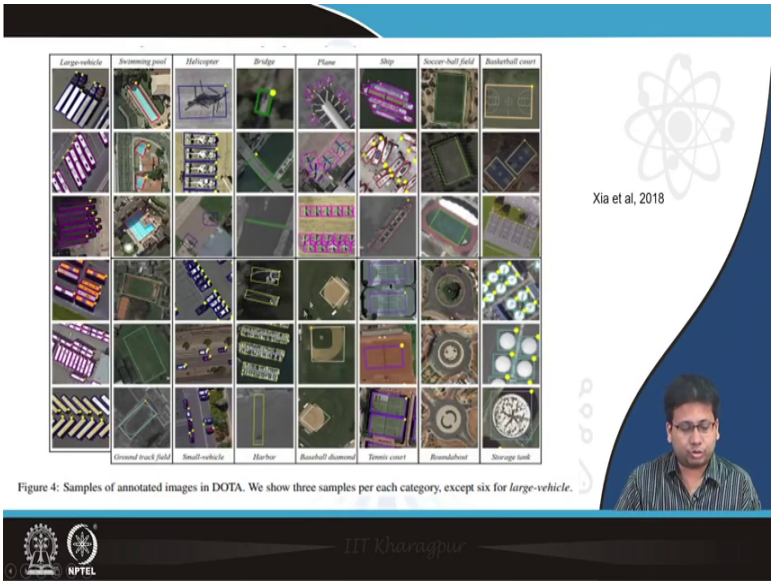
To advance object detection research in Earth Vision also known as Earth Observation and Remote Sensing, we introduced a large scale Dataset for Object Detection in Aerial images or DOTA. In this end, we collect 2806 aerial images from different sensors and platforms. Each

image is of size about 4000×4000 pixels and contains objects exhibiting a wide variety of scales, orientations and shapes. These DOTA images are then annotated by experts in aerial image interpretation using 15 common object categories. The fully annotated DOTA images contain 188282 instances, each of which is labelled by an arbitrary quadrilateral.

To build a baseline for object detection in Earth Vision, we evaluate state of the earth object detection algorithms on DOTA. Experiments demonstrate that DOTA well represents real Earth Vision applications and are quite challenging. So, like these are some of the examples from this DOTA data set. So, like as you can see this is the scene of a something like a jetty where a large number of boats are parked in close like closely to with each other.

And, you can as you can see around each of the boats they have put a bounding box; so, that like to indicate that this is the precise location. So, the in this case the object class is boat, but then there are many instances of the same class, similarly in the other images also.

(Refer Slide Time: 03:51)



So, like now the 2806 images there are. Now, as I said like this number 2806 is actually quite small. In fact, very small as far as normal computer vision tasks go. Like for us to develop a fully functional deep learning framework which is capable of doing say object detection on natural images and something like that, we usually need millions of images. But, in this case we do not

have so many, we have only couple of or maybe just about 3000 images, but each of them are very high resolution.

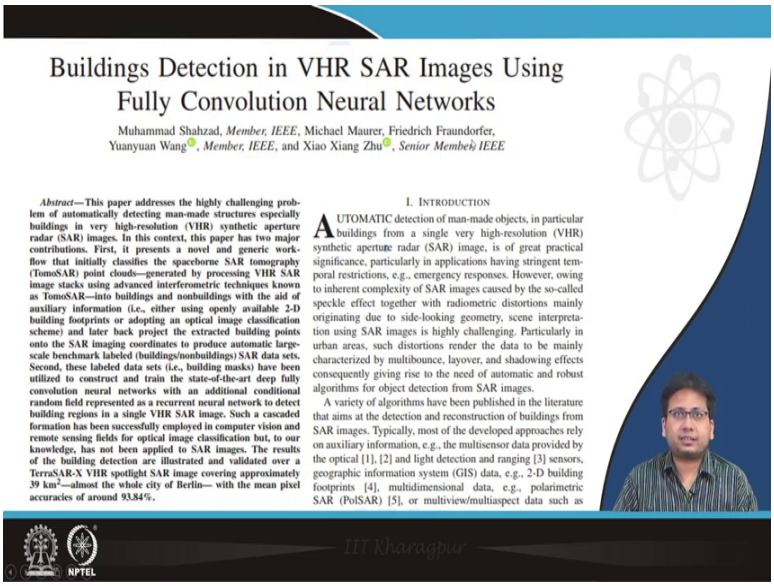
So, what can be done is we can break up each image into multiple sub image and then do the whatever we want to do on them. So, that will increase the number of images like and like while we are like we can also do the labelling on each of them. And so, like these like as they have mentioned in this data set there are 15 object categories. So, some examples of that they have shown like tennis court, like the harbour, small vehicles, ground track, etcetera.

So, like these and in each of the cases as you can see they are these pink or green or yellow quadrilaterals, these actually show the local precise locations of the corresponding objects within the those images. So, this date it is an useful data set on which many of the like object detection models can be trained. Now, like we know or we or most of us have heard that they are like it is often very difficult to completely train a neural network from the scratch on any data set.

Because, that requires a very high amount of data, I mean I a proper neural deep learning framework model will have like millions of parameters. So, to estimate them we will also need millions of like training images which is difficult. So, what people do is they build pre trained models or they use pre trained models which have already been trained on some other data set. And, then on the data set which they are interested in they retrain only some parts only a may be only a few parameters and so on.

So, effectively they are they take a general model and then they personalize it on the particular thing that they are interested in, on the particular data set that they are interested in. So, this DOTA data set the by this provides a like a source on which a model like this can be pre trained and then that pre trained model can be used on specific remote sensing related problems.

(Refer Slide Time: 06:50)

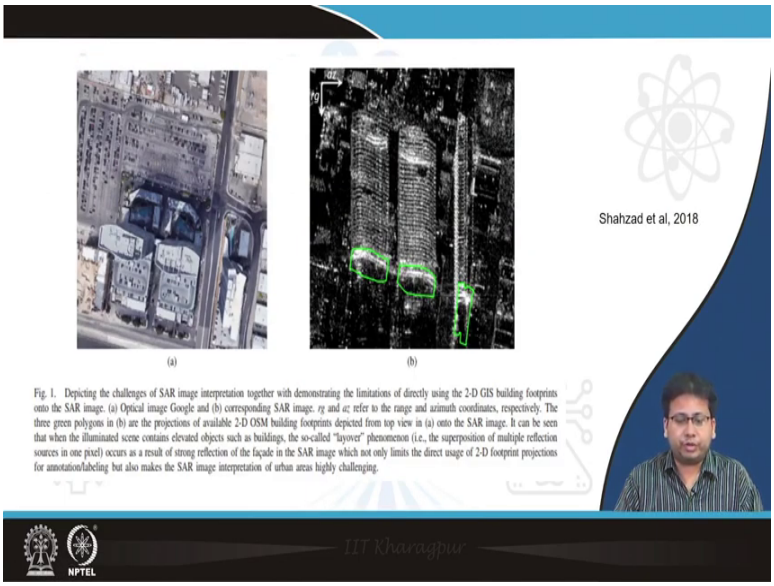


So, building detections in VHR SAR images using fully convolutional neural networks. So, VHR means Very High Resolution and SAR is synthetic aperture radius. So, that so, sorry Synthetic Aperture Radar technology which we of related to active sensing which we have already discussed in the previous lecture. So, on the images captured in using this technology, they want to detect buildings. This paper addresses the highly challenging problem of automatically detecting manmade structures, especially buildings in very high resolution, synthetic aperture radar images.

In this context, this paper has two major contributions. First, it presents a novel and generic workflow that initially classifies the spaceborne SAR tomography or point clouds generated by processing VHR SAR image stacks etcetera into buildings and non-buildings with the aid of auxiliary information, that is which either using openly available 2D building footprints or adopting an optical image classification scheme.

And, later back project the extracted building points onto the SAR imaging coordinates to build or to produce automatic large scale benchmark labeled or building non-building SAR data sets. So, as I said the images which you get using some of the sensing technology like SAR, they may be different from the human that is the optical images which we are familiar.

(Refer Slide Time: 08:28)



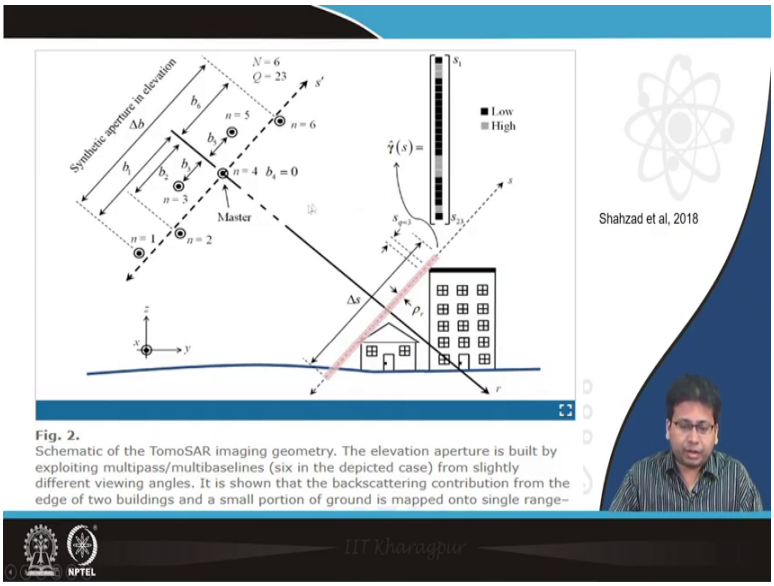
Say for example, this is an optical image which is taken like within the which is captured within the visual visible range by something like a drone or something like that. So, these are the different buildings. So, as by looking at it you can understand that here is the top view of a building. This suggests the top view of another building and like these also may be small buildings and so on.

These seem to be trees etcetera. These are or the these also may be and some another building like that. But, when we are talking about a synthetic aperture radar image like this the SAR image, then this is the response you get. Now, by looking at this you will probably not understand that there are any buildings at, all I mean nothing looks like a building here. So, what they need to do is they need to identify buildings in this kind of imagery which is of course, like very difficult.

Because, we like I mean even the training it is for the training we need labels that itself will be difficult, because a human setting with these kind of an image will not be able to understand where the building is. So, they want auxiliary information like an optical images like this and then within the optical image, they can actually like a human expert can identify the locations of the images. And, those like those pixels which are known from the optical image to contain a building, they are mapped to the corresponding pixels in the SAR image.

So, that we know that these like in the SAR regime, these are the representations of buildings. So, based on these kinds of representations we have to build or we have to find a the other buildings in similar images.

(Refer Slide Time: 10:22)



Now, when an SAR image is being captured there are of let us say of a building, then like there are many factors which like which determine what will be the nature of the response to the building. So, like as you can understand like sa as SAR is an active is an active sensing technology. What it does is like there is something like a laser which sends out a rays and then like this and then the rays are reflected back.

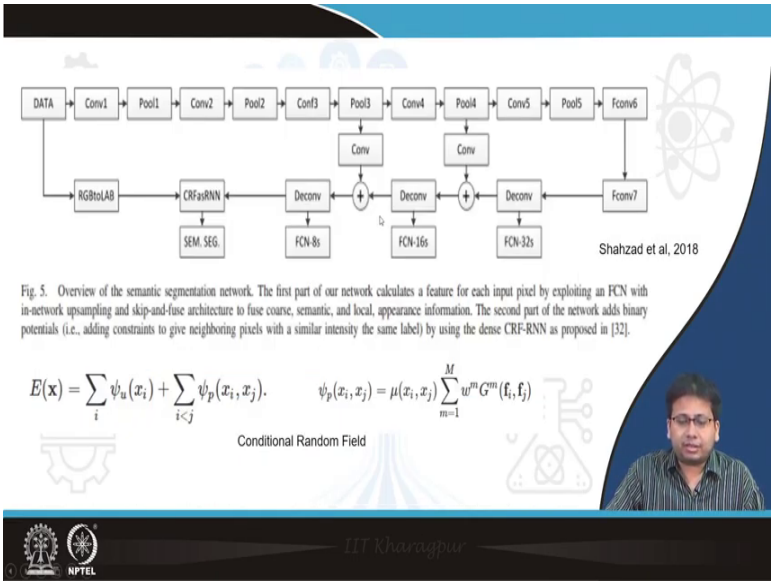
And, based on the reflection they try to understand like basically they measure the intensities of the reflection and so on. Now, how much of the sent or emitted energy is going to be radiate or reflected back to the to the sensor, that depends on so, many things like the angle at which it is sent. And, and this is the various things like that as is as you can understand from this image. So, like here is a building and here is a building.

So, like had the sensor been directly on top of them, they would have responded in one way. I mean then the emitted signal would have come down and then just reflected back, in this case also it would have been come down and reflected like that. But, then again this one as you can

see this has a inclined roof. So, like if it like then again the way that signal is going to be reflected back, we know is going to follow the laws of reflection. So, like we may not get like the amount of energy that is reflected back may not match the how the same energy, that is it is being emitted for the structure for the shape of the roof and so on.

Then, like as you can see in this case the angle, the relative angle of the sensors with respect to the buildings it might be something like that. So, the emitted rays instead of hitting the roof may hit some a strange part of the building like that in which and that too at an angle. As a result of which only a small part of it will be reflected back. These are some of the challenges why detection of buildings or for that matter anything, any other things using this kind of a SAR image is difficult.

(Refer Slide Time: 12:44)



Now, in this particular paper what they have done as I said like they have like developed the labeled examples with the help of auxiliary information related to where the buildings might be located. So, these auxiliary information may be the list of building coordinates that they know or by comparing with an optical image like this.

So, now once they have got the training data, next they want to build something like a neural network like which will be able to do the building classification task or I mean basically they

want to classify every pixel as building or non-building. So, for that they use these kind of a like a deep neural network to represent the. So, these like these convolution, pooling convolution, pooling all these layers they basically are used for representing the pixel level information from the SAR image. And, then once that is done the those pixel level information are then refined using what is known as a CRF, a Conditional Random Field to like produce the building, non-building classification.

Now, what is like the and that is basically how the detection takes place, that is to say each of the pixels is like is finally, classified as is either it is a building pixel or a non-building pixel. And, like so, the like we can say that a building has been detected if there is a like at any particular part of the image, there is a like a box of building pixels like that as classify that is pixels that have been classified as buildings.

Now, these there is they use this concept of CRF: Conditional Random Fields to ensure something like smoothness of the labels. That is we do not like if that is common sense suggest that if this pixel is building, then it is not likely that the adjacent pixel will be non-build or the adjacent pixels are going to be non-building. Rather, all the like the since building is a contiguous object, if there is a building pixel here its surrounding pixel should also be building. If there is a non-building pixel here, its surrounding pixels should also be building. So, these kind of consistency, spatial consistency is like encouraged by this kind of a conditional random field.

So, if you remember in the past lectures, we had talked about enforcing spatial coherence using Markov random fields, where we defined some potential functions for the same purpose. In this case also basically the same thing is being done. So, a conditional random field is a also a graphical model similar to a Markov random field which is also used to like represent a joint distribution. So, in this case the and like Markov random field, the conditional random field also has some potential functions.

So, those so, this is these are the potential functions. These are the pixel wise potential functions and these are the potential functions on pairs of adjacent pixels. And, these potential functions are defined in a in a way using these kinds of Gaussian kernels that operate on the pixel-level

features. So, as to ensure that the corresponding pixels get the same labels as that is either both should be buildings or both should be non-buildings and so on.

(Refer Slide Time: 16:23)

So, they this is how they like that paper, it carries out the detection of buildings from SAR images. Next coming to another paper. So, here the task is to develop a framework for large scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural network.

So, instead of like specific buildings, they want to detect human settlements which includes buildings like ah, but can, but in general indicates like more like a something like a city, town, village, this kind of thing. So, human settlement extent information is a valuable indicator of worldwide urbanization as well as the resulting human pressure on the natural environment. Therefore, mapping HSE is critical for various environmental issues at local, regional and even global scales.

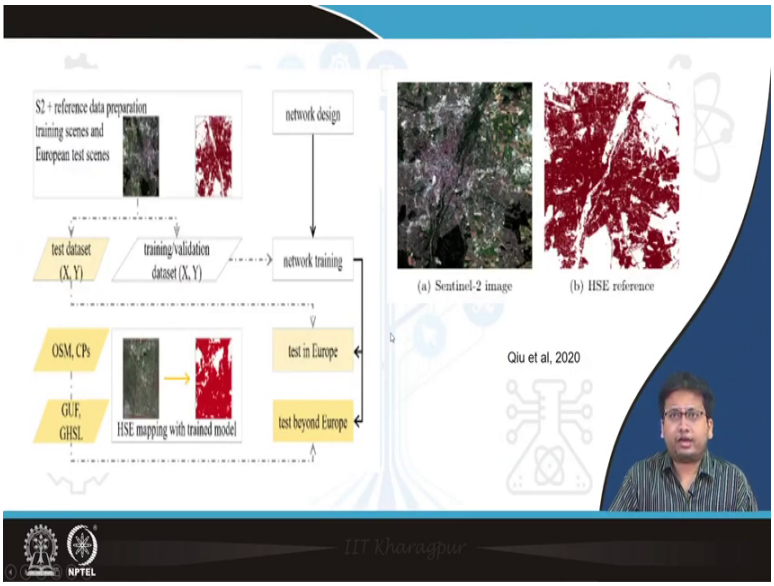
This paper represents a this paper presents a deep learning based framework to automatically map human settlement extent from multi-spectral Sentinel-2 data using regionally available geo products as training levels. A straightforward, simple, yet effective fully convolutional neural network architecture sent to HSE is implemented as an example for semantic segmentation

within the framework. The framework is validated against both manually labeled checking points distributed evenly over the test areas and the open street map building layer.

The HSE mapping results were extensively compared to various baseline products in order to thoroughly evaluate the effectiveness of the proposed HSE mapping framework. The HSE mapping framework is consistently demonstrated over 10 representative areas of the world. We also present one regional scale and one country scale HSE mapping example from our framework to show the its potential for upscaling.

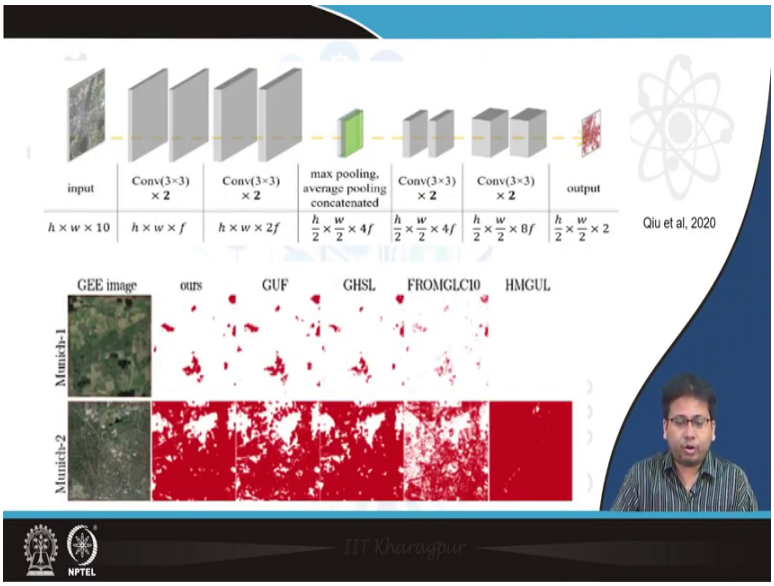
The results of this study contribute to the generalization of the applicability of CNN-based approaches for a large scale urban mapping to cases where no up to date and accurate ground truth is available, as well as the subsequent monitor of global urbanization. So, here several new concepts are being used. The first of all like this is based on like as they mentioned this is on the multi-spectral imagery from the Sentinel-2.

(Refer Slide Time: 19:02)



So, as I had mentioned earlier, multi spectral image is such an image where like the images are cap, there are multiple images captured at different wavelengths which are widely spaced all over the spectrum. The number of images is captured is not much, typically say 3 to 15. But, their wavelengths are very different from each other.

(Refer Slide Time: 19:30)



So, the other important concept which they use in this paper is that of domain adaptation. That is they have trained it on one part of the world and they have used it on other parts of the world, that is, this shows the generalizability of the neural network which they have built. The reason for doing, this the reason why this is significant is there are some parts of the world especially the developed countries like US, the Europe and so on.

Basically, what is known as the global north, where like these kinds of aerial or satellite images might be obtained very easily and it might be possible to label them also. I mean the satellite images of course, can be taken from anywhere over the globe because, it does not matter whether it is a developed region or a undeveloped region. Because, the satellites they cover the whole world, but like the thing is we need to train those the model by using the ground truth labels.

Now, the we can hope to get accurate ground truth labels only in those places where we have detailed maps and so on. So, that is possible only in the developed countries of the global north.

Now, like once the model has been developed there, then it can be easily exported to other places or less developed places where it is may not be possible to retrain the model. But, the if the this work shows that the model should continue to work even in those cases because, it the according

to the claim the claim is that the model is the architecture is general enough. So, the general out outline of this approach is as follows.

So, first they have the reference data preparations, they have the training scenes and the ground truth labels, that is these red indicates the places where the humans are actually settled. And, the these white they may indicate rivers or the ones near the edges they may indicate open fields or pasture lands and so on. So, this is the image captured and this is the ground truth level. So, the all these things are fed to the these pairs are fed to the network which they have designed for the training purpose.

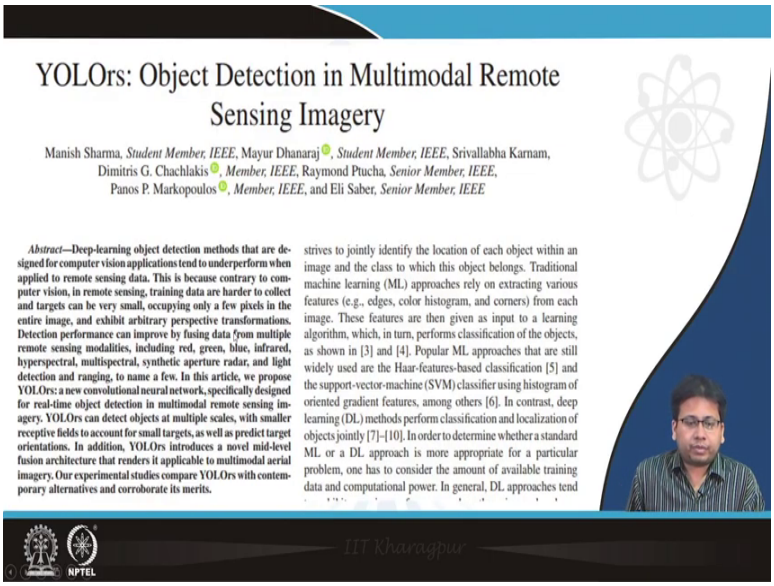
So, then the network gets trained and then so, basically the it is now able to map the HSE, I mean the multi spectral image from any region to the HSE map, that is the human settlement map the and that is while the model is first trained over some cities or some regions in Europe. Then it so, it is tested first it is tested over other in the same maybe some other European cities and then it is tested beyond Europe also, in some other places where maybe they have the data for testing. So, this is what the neural network looks likes.

So, as you can see again there is an its a basically it is a convolutional neural network only, that as architectural architecture wise there is no innovation. This is the input image, its like as usual it undergoes a series of convolutions and poolings. And finally, they get this kind of an output where at every pixel contains the information whether it corresponds to HSE or not.

So, this is the like this is an image which may they have obtained from Google Earth Engine. And, these are corresponding to that these are the mappings of the human settlement which they have found. And so, like they have evaluated it against various other approaches for the same thing that is HSE mapping. And, they show that they their approach is better like compared to many of the others which are like some of them either overestimate or underestimate the extents of the human settlement.

Especially, if they like if they are scattered rather than being dense. So however, the proposed method is actually according to their evaluation is able to do that to do the identification of the HSE in a very accurate way.

(Refer Slide Time: 23:46)



And, then we come to the last paper which we are going to discuss in this lecture. So, YOLO we have discussed earlier also the You Only Look Once. The framework for object detection in natural images or the optical images. In this case, they are proposing YOLOrs that is with YOLO specifically for remote sensing imagery. Deep learning object detection methods that are designed for computer vision applications, tend to underperform when applied to remote sensing data.

This is because contrary to computer vision, in remote sensing, training data are harder to collect and targets can be quite very small occupying only a few pixels in the entire image and exhibit arbitrary perspective transformations. Detection performance can improve by fusing data from multiple remote sensing modalities, including red, green, blue, infrared, hyperspectral, multispectral, synthetic aperture radar and light detection and ranging to name a few.

Ah In this article, we propose YOLOrs a new convolutional neural network specifically designed for real-time object detection in multimodal remote sensing imagery. YOLOrs can detect objects at multiple scales with smaller receptive fields to account for small targets as well as predict target orientations. In addition, YOLOrs introduces a novel mid-level fusion architecture that renders it applicable to multimodal aerial imagery. Our experimental studies compare YOLOrs with contemporary alternatives and corroborate its merits.

(Refer Slide Time: 25:44)

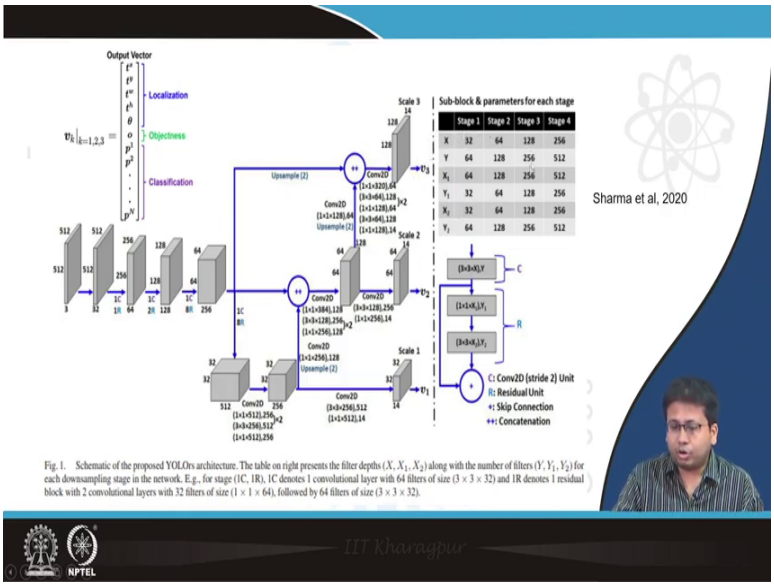


Fig. 1. Schematic of the proposed YOLOrs architecture. The table on right presents the filter depths (X, X_1, X_2) along with the number of filters (Y, Y_1, Y_2) for each downsampling stage in the network. E.g., for stage (1C, 1R), 1C denotes 1 convolutional layer with 64 filters of size ($3 \times 3 \times 32$) and 1R denotes 1 residual block with 2 convolutional layers with 32 filters of size ($1 \times 1 \times 64$), followed by 64 filters of size ($3 \times 3 \times 32$).

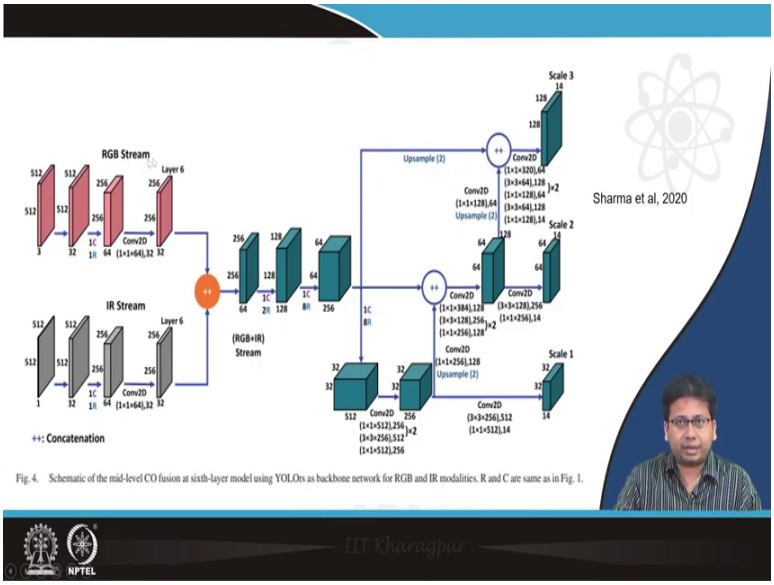
So, the important the most important concept here is that of multimodal remote sensing imagery that is instead of depending on one source of image or one technology of this remote sensing imagery, they are using multiple sources. And, and then they are doing the detection on the of the objects at multiple scales. And, like is like as mean discover discussed for YOLO also, the what the YOLO does is it builds a hierarchy of representations at multiple scales and etcetera.

So, that at each scale the different sub-parts of the image are classified as object or it has belonging to the target object class or not. So, in this case what like the YOLO in this case of the YOLO or YOLOrs. So, this is like what the output vector will look like. So, we first of all it will have classification that is indicating whether the different like we can say that like it will indicate whether the different object classes are present or not, or if so, what is the probabilities of each of them.

And, then there is localization which suggests like the for the for every instance of the object class what are its exact locations, that is, these the exact coordinates in the image. And, then there is also it gives something called o, the objectness score. So, this is measured at every pixel where like or not I mean not at every pixel, but at every region where they try to like where this objectness score indicates whether that region it corresponds to a particular object or not. So, like this is the broad representation. So, like once again it is a CNN only.

So, there are these convolutional and pooling layers and then the whole thing is like is broken up into multiple scales. And, then at each scale the convolution is done separately to get the pixel level classification at each of the different scales. And, then like all the scale level outputs I mean the outputs at each of the different scales or the subblocks etcetera, they are somehow combined together to come up with the final layer.

(Refer Slide Time: 28:09)



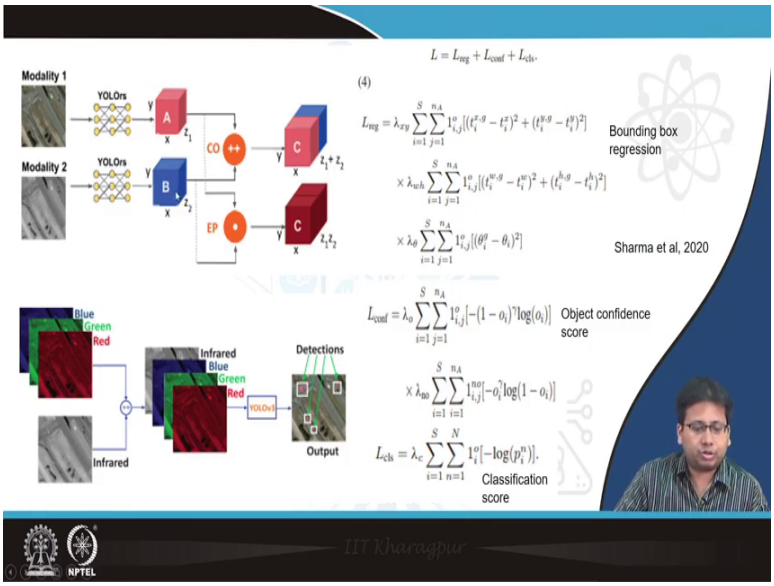
And so, when we do are doing this in a multimodal setting as they mentioned. So, let us say that there is a there are two sources of image. There is a RGB image, that is the normal optical range visual range image and an infrared image like that. Both are of the same both images are taken over the same region and the images are also of the same size. In this case both are of 512×512.

So, both of the images are first represented independently by a convolutional neural network. So, like neural you can say for both the images neural features are developed, then there is this concatenation layer where the image the image features from both sources, they are somehow fused together. And, then what we have is a combined representation.

So, we will discuss a bit more of this concept, when in the next lecture when we talk about image fusion. Now, this combined representation is again passed through several convolutional layers and so on. And, then after that this combined image is again subjected to the multi-scale analysis

that we discussed in the previous case also. And, like at each of the scales the like object detection in the pixel-wise classification, that is object or non-object is carried out. And then finally, they are merged together to form the pixel-level output.

(Refer Slide Time: 29:37)



So, like roughly it looks somewhat like this. So, this is the image from 1 modality, let us say RGB and this is from the second one, let us say modality 2. Now, both of them are passed through some kind of YOLO framework. And then like their representations, they are somehow concatenated together and then like we get like we can get the thing. So, this is one approach where the YOLO is done like separately on each of the approaches, on each of the modalities.

So, this is you can say it is something like a baseline approach. You do not fuse the things right away, you on each of the images or each source of images you do the object detection separately and then combine the results the or, but that is not what this paper is proposing. This paper is proposing the another thing, it is saying that you have images from both sources you somehow build a combined representation of them. You bring them to something like a common space, a common ground by doing the neural operations like these convolutions.

And, then once you have that kind of a combined image representation on that one you apply the YOLO and get the detections that is for at each pixel, you get the that is you classify it as either

belonging to the object class or not. So, this is the like this is the approach which they have proposed and they have claimed, that this is better than the previous than this approach. And, like whenever we are training any neural or any machine learning algorithm we of course, need a loss function.

So, in this case they have three parts in the loss. One is the regression loss, the confidence object confidence loss and the classification score. So, the these regression loss, this is related to the bounding box. So, like we know that for whenever there is a object class, it should have like or any object instance it must have a bounding box around it. And so, and as discussed here also like for every instance, they show the or they predict the exact coordinates of the bounding box.

So, in the loss function that the aim is to compare the predicted coordinates of the bounding boxes with the actual coordinates of the bounding boxes of that particular object. And so, this is this can basically be considered as a regression problem so, that gives you the regression loss. Then, there is the object confidence loss. So, suppose a bounding box has been found somewhere.

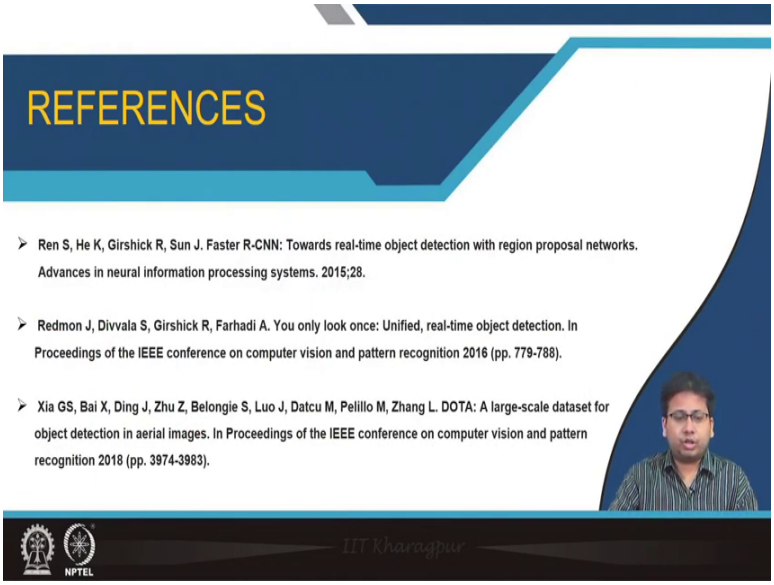
So, now what is the likelihood that it actually contains some concrete object or that is its not that its that bounding box, it like it is on the background or it or its like overlaps; like part of it is on an object, part of it is on the background something like that. So, for to measure that thing they have something known as the object confidence score. The probability that these bounding box contains an object.

So, that is somehow measure that I mean of course, that cannot be directly measured, but from what the response of the neural network this o variable is created and this is supposed to represent the objectness score. So, the based on these they build the object confidence and finally, the classification. So, I mean ultimately the each of the bounding boxes, we have to like suppose I have shown you a bounding box.

I must also predict what is contained in that bounding box, what kind of object is there. Like is it a building or is it a car or is it a cat or is it a dog or what is there. So, that is the classification score. So, there are these three things; the regression, the confidence and the classification. Like all these are concepts are used to define a loss function. And, based on and this the neural

network that is being discussed here, that is the YOLOrs that is trained to minimize the loss function discussed in this way.

(Refer Slide Time: 33:54)



So, these are the references that we discussed. The first two are related to R-CNN and YOLO, that is the original papers that we discussed in the previous lecture and yeah. So, that brings us to the end of this lecture. So, we will continue our discussion on like computer vision on remote sensing imagery in the coming lecture.

So, till then bye.