Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-78

Then you can also have a situation like this which is actually semi-supervised and this is like monocular. See till now what, I mean I have kind of tried to put them in some sequence so that you understand. So the first two as I said are kind of block matching and therefore it is only a part of, it is not end to end. The other paper that I showed was bringing in the context but then again it is not end to end. Even third one that we talked about was unsupervised that is actually end to end and it is also unsupervised but then that is for a stereo setup. Now this other paper that is here that is about monocular depth.

That means you do not have really a stereo pair. So how do you actually do this? And by semi-supervised what they mean is there is a depth map that is coming from a different sensor which is let us say a LIDAR. In this case I think it is basically a 3D laser. So what this gives you is kind of a sparse sort of a null depth map.

So here is a scene for you and then if you had a dense depth map, this is how ideally it should look like but then what you have is a sparse sort of a null depth map and therefore it is called actually semi-supervised. So the points where you have a depth map or the information especially that is coming from the other sensor, that part you can use as a ground truth and then for the rest of the portions you do not have the ground truth. And therefore again you have to do this kind of a left-right consistency in order to check are you able to match the images correctly or not. But then this is like I say monocular case and then the other one that I also kind of wanted to briefly touch upon is this wide context learning. So this again coming back to stereo matching but then you see this one, this network that they have thrown in between, it is called a wide context learning.

And what does wide context again means is that the earlier case that you saw was a multi-scale kind of a situation where you were down sampling and then trying to sort of increase this receptive field. This is a direct way of rate increasing receptive field. So you must have heard about SPP, spatial this one, pyramid pooling. We did not do it, I mean the codes but then it simply means actually it is used more for the case wherein whatever be the input size if you want to finally reduce them to one single sort of size, that is where SPP comes. But in this context why we think of it as a wide context it is like this AP is

actually average pooling and this 8 by 8 what this means is that you are actually dividing whatever be the feature map that is coming in, you are actually breaking it down into this 8 cross 8 regions and you are trying to map, so you are doing average pooling on let us say each of those blocks.

And the idea is that, so for example, the information rate that is coming from this 8 by 8 is like I mean so if you look at the averaging that is going on it, you are averaging over much larger region then when you go to 64 cross 64 that means you are averaging over a smaller region perhaps. But then irrespective of what the feature map is having but the point is that in one case you get to do this average pooling over a larger region, in another case you get to do this average pooling over a small region and the idea is that the context rate is what is then it entering into the picture. And there is still no clear insights into how this helps for example, how this helps you counter noise, how this helps you counter illumination and all because that kind of an explainability is still lacking, people have been only able to show results, so it is like intuitive, which is also the reason why the other way to do it is using a dilated convolution, they use both right, you guys know about this is a dilated convolution right. Dilated simply means that right you instead of a normal convolution right I mean see for example, I mean if I did a normal convolution what I would do let us say it is a 3 cross 3 then I will kind of you know so if this is my input image right I will just look at these locations and then this right I just whatever is that value I copy here. Now in a kind of a dilated convolution right you tell the kind of a dilation rate, so for example, right if you have a certain rate then what could happen is then instead of sampling this image like this right you would actually sample it you know on a kind of so you will still have a 3 cross 3 kernel only, but then but then right the it is kind of spread out right.

So, you have so instead of sampling these adjacent locations we are sampling with some gaps in between right, it is also called the at-risk convolution I am sure that many of you have probably read it somewhere. So, these are all different types of convolutions and one of them is this is also called dilated or I think the other word used for that is at-risk convolution and this actually helps you kind of get a kind of you know larger field I mean without actually involving more unknowns right there is still 3 cross 3 only I still have those 9 parameters imagine I mean if I wanted a 5 cross 5 right then I would have had to have one 3 cross 3 followed by like one more get a 3 cross 3 if I was doing a regular convolution right that is how I will get a 5 cross 5 receptive field. Now in this case right of course you are sacrificing a little bit in the sense that in the sense that you know in between right you are you are not you are not caring about it, but that is the way it is right at-risk convolution works that way right and these guys have tried to incorporate wide context right. So, this context business and wide context right it is coming through the spatial pyramid pooling and this is a dilated convolution. Yeah, then the left right

consistency right I have already talked about, but then right the only sort of you know a difference in this paper is right they use they use one of the images that is let us say the left image do a prediction of both the right the left disparity as well as the right disparity.

And then and then and then and then what they will do is they will take the take the left image use the use the right left sort of you know no use the right image right image use the left disparity produce the produce an estimate of ihel then they will use the left image use use the right disparity along with it to actually produce i r hat and they will compare at i r and i l because i r and i l you know how they are. So, this left right consistency is like over both left and right I mean you walk once through a right disparity map once through a left disparity map. So, i l will be acted upon by d r i r will be acted upon d l and if you acted upon then you should actually get back your original pair because that is all that you have. So, again it is unsupervised right and this is another way to do and then this is actually deep m v s. So, this is like the multi multi view stereo thing right that we that we did and this actually goes back and uses all that plane sweep stereo and all right.

So, the idea is like this. So, here right if you look at this plane sweep volume right this is like saying that for you see one depth right you will get kind of one volume for another depth right you will get another volume for another depth right you will get one volume. So, depth or kind of you know a disparity. So, I think they use like 100 of them right 0 kind of you know disparity then 99th disparity. So, it means they have got they have sliced the 3 d c into 100 planes.

So, what I am trying to say is right many of these methods do go back and use the things that are actually classical in nature, but how they kind of put that in and how they in what form they use it how they use it is what is what kind of rate differs from one paper to another. So, this plane sweep volume right we know right what it actually means and then what they do is they have they have a kind of rate of photo consistency you know right sort of a network what this actually does is that is that right. So, it actually creates a creates a creates a 400 400 channel. So, channel sort of an output and then this is then passed through this unit right and this unit also gets the semantic information that is why that is why right I mean. So, this context right that we are talking about here the context is coming through you know a different way in which in which it is coming through a v g g net right and that input is also sent to this unit along with this along with this kind of along with this with this feature volume right that you have and outcomes kind of a you know 800 800 feature layer and this and this right and this they do for do for every this one a reference a reference view.

See if you look at this right if you if you actually notice it for one volume rate you get actually a 4 channel output right and then you have 100s 100 of them which is the reason

why you have this 400 right and then right then this 400 goes into this network and then outcomes and you know 800 channel layer and this they still call this intra volume feature aggregation right because everything is happening is happening with respect to with respect to one reference frame and then it everything else is being warped at let us say various depth planes then they use what is called inter volume. So, what this inter volume means is that now you change instead of using let us say the first reference image now use a second as a reference now you do you know right repeat the entire operation then you get another sort of an 800 feature vector then you then you kind of do it for the do it for the next image. So, so this input is like how as many views as you have right I mean here here right it is like it is like as many as a disparity levels as you have here it is like as many views as you have and all those are sent into some max pooling again right that is why it is called it is called inter volume feature aggregation and then it is supervised by the way right it does not unsupervised at the output you predict a disparity map. So, for example, I mean so you have you have this 100 values right so it is like a it is like a it is like a you know 100 valued sort of a disparity map and you again compare it with actually with a ground truth and the ground truth here I believe is synthesized right so it is called MBS synth. So, this is actually completely synthetic data set but they claim that they have been able to introduce lots of things into it like color changes whatever I said illumination all of that is sitting there and then they say that right once you train this network on this what is this multi view stereo synth data set right then they show that they show that right you can take it to a to a you know a different data set and you can see right for example, if you see this building right and and right I mean you know outcomes outcomes outcomes you know 3D structure like that then from here to here right you can still see this kind of kind of building right coming up in the front and whatever right and they also also also compare with what is called you know cold map is a standard kind of structure from motion kind of a traditional method.

So, they compare it and they show that probably right with deep deep MBS you got much less noise and it is far more accurate and so on right. Then the last one that I talked about what I wanted to talk about was structure from motion all the earlier ones were all on stereo that you can also ask about is the structure from motion right. So, what if you what you have had a which you had a monocular video. So, here the idea is that right you have lots of unlabeled video clips right what that means is that I mean you do not I mean you do not you do not know the kind of structure or anything you just have you just have several video clips taken from a camera and what you do is you know you actually build you know two of these networks right one is called called a depth CNN another is called is actually called you know a post CNN. So, the idea is that how do you do unsupervised learning of depth and ego motion by the way I write this word ego motion I should have used it I think I never used this ego motion because ego motion means the camera motion that is what that is what basically you know ego motion means and the idea is that right

just like we do right in structure from motion what would we have done right we would have said calculate for me all the all the camera poses as well as the structure of the CNN that is what we do in SFM right that is exactly what is being asked here what is the ego motion and what is the depth and the way it works is as follows right.

So, you have sort of you know a reference view that you can pick and you actually build a depth CNN and this completely unsupervised okay so what you do is you kind of have a depth CNN let us say right you know initially it is supposed to it is supposed to give out a depth map initially of course right it will not be able to do it will give you something right and that is that is your D right that is sort of you know a depth map that you have with you then what you do is you give your nearby nearby views. Now if I if I remember right right what you do is they actually train with 3 frames that means you have like you know a reference frame which is TT they have TT - 1 then they have C right TT + 1 okay and they and they kind of and they ask for the pose between this and this and kind of read this and this that is called that is called a pose network. So, this pose network is supposed to give you let us say R1 T1 and then it is supposed to give you R2 comma C T2 okay and then the and then the idea is that right you go back and and do the see now this is not like left right consistency right this is not like stereo match so you do not call it a left right consistency what you do is you still have to look for this called view consistency okay the even difference between the earlier ones where we did stereo we kept on saying left right left right but it is not left right right because this is like a view. So, I have one view from here another view from here I am moving and then I am getting different different views and I am looking for view consistency and the view consistency right if you go back what is that equation I do not exactly know maybe you can remind what is that X ~ dash right we had when I was talking about view synthesis you guys remember it was some some K inverse and I think there was some R T which was the pose and then you have a vector right Z and then you had another K inverse I do not know which one was KL and which one was KR I think this was I think this was KR KL and this was KR sorry and then you had a KR Z KR inverse and then you had an X~and then 1 it is something like this we had right if you guys going to say if you guys recollect okay this is what I had said people use for active view synthesis right and this is exactly the equation that goes in here because the view synthesis you cannot just do it in air right I mean you have to do some kind of a view therefore you have the R and T right and then Z Z is is is right is from here so you have a depth net which gives you the Z you have the R and T which comes from which comes from the post net and the intrinsic it is supposed to be known and typically you do not have KL and KR this is the same camera it is just 1 K right and the idea is that right you do the view synthesis and what is the view that you have you already have these views with you right so what do you do when you kind of take this T T which is a reference view you warp it you warp it so as to match your T T - 1 view and T T + 1 view using this D and right and this Z sorry this D and this R and T the post net and and the right if there is

an error then you sort of go back right then you kind of keep training this network and you can show it right tons and tons of video clips the nice thing about this network is you do not need anything right you just have to have this have this right videos taken right tons of them keep on keep on giving them and ask for ask for a ask for a view consistency check it simply means that right it needs to keep on creating the synthesizing the views and then at the at the end of it right it is completely trained. Now this network you can actually use it in use it in several ways right sometimes you can you can even you can throw away all of this right you may not even need this and I can simply use this depth CNN what that means is a monocular depth that means I just give I just give you give this network an image and then and then and then it knows to it knows to give me a depth map for that right I do not need anything else I can just give it a single after training right after training I can I can either use this whole thing as it is that means that I can give frames like you know 3 frames and then and then at one target view and then at the output it will give me a depth map corresponding to the target view and then it will give me r and t which are the which are the other poses if I need them.

Other way to use it is throw away everything use simply a depth CNN wherein wherein any image that you input because this depth CNN has figured out how to kind of compute a depth map given because because it only sees that image right the input image and therefore it can actually give you a depth map you can also use the post net separately if you wish that means that if I have if I have 3 frames and I and I and I write and I want to know what is the relation between them right it can it can I because it just needs those frames you can train it with as many frames as you want right depends upon how you have trained it if you have trained it with 3 frames and for those 3 frames it will give the give the you know mutual poses this can all be useful depending upon it what kind of a problem you are dealing with for example that if you are doing some kind of a video super resolution and so on and if you want to know R and T and you are not interested in other things then you can probably use the post net alone. So all of that right is entirely sort of out in the open right how you want to do that and this was actually a 2017 paper right so anyway there are there are say tons and tons of papers ideas not to go through all of them but to sort of give you a rough sketch about what is what is deep networks kind of do and you should be able to read them up right now that you have enough background you can just go read them up and you will be able to understand what they are doing and why they are doing there are few other things like this cost volume and cost aggregation and all right which I have not talked about within these networks that I mean you will have to read the papers I mean there is more kind of you know detail right out there but yeah if you feel that certain things you want to understand better I have given you the paper titles just go there and read them up.