3D Reconstruction and Estimation from Single-view 2D Image by Deep Learning – A Survey

Yongfeng Shan
School of Computer Science
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
yongfeng.shan@student.uts.edu.au

Christy Jie Liang
School of Computer Science
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
jie.liang@uts.edu.au

Min Xu
School of Electrical and Data
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
min.xu@uts.edu.au

Abstract—3D Object Reconstruction from a single-view 2D image has become a promising research field. However, it remains a crucial and unsolved core issue in AI and Computer Vision research. Many scholars think it is the future of Artificial Intelligence and is well deserved at the irreplaceable heart of future AI research. In this paper, the research history of 3D object reconstruction is introduced, and the current state-of-the-art research methods and most novel results are investigated and discussed. A prediction for the best research methods and reconstruction model for this field is made. This paper also provided the essential clues and trends in 2D images or scenes inverse to 3D sceneries by 3D reconstruction. This interdisciplinary research area requires the researchers to have rich knowledge in but not limited to Computer graphics (Such as OpenGL), deep learning, Computer vision (like OpenCV), and the neurocognitive logics and principles of the cerebral cortex.

Keywords—Single-view image 3D reconstruction, 3D Scene Semantic Understanding, 2D Image 3D reconstruction, Deep Learning, 3D CNN

I. Introduction

Artificial Intelligence(AI) has become the centre of Scientific research, among which the CNN (Convolution Neural Network) [1] play the most pivotal position. Furthermore, based on CNN networks, the Deep learning area has brought us many outstanding performances in computer vision research and people's everyday life [2], [3], [4], [5], [6]. All the activities and tasks, such as Object Detection, Semantic Segmentation, Computer Games Aid, Pictures and Video recognition and classification, have progressed to a very advanced level compared with ever before, which are primarily sponsored and benefited by the development of deep learning.

However, the prevalent use of Convolutional Neural Networks (CNNs) [1], such as ResNet [2], trained predominantly on 2D image datasets like ImageNet [7] and MS COCO [8], imposes a significant limitation in object detection tasks: the loss of depth and spatial context when translating the three-dimensional world to a two-dimensional image plane. This limitation is particularly acute in applications requiring a deep understanding of 3D spaces, such as autonomous driving, robotics vision, Metaverse, and Medical Imaging, where accurate depth perception is crucial. While these CNNs excel in parsing detailed pixel information, they inherently need more capacity to process and interpret the z-axis data indicative of depth, resulting in a critical disparity between the machine's perception and the nuanced spatial awareness akin to human vision. Furthermore, the low resolution of the point clouds and the prohibitive costs of high-resolution LiDAR systems [9], [10] compound these challenges, underscoring the urgent need for advanced deep learning techniques to reconcile the 2D training paradigm within the 3D operational context.

To tackle the above issues, how to make the 3D shape and semantic reconstruction and estimating from a single image appears to be the critical and essential part. However, the 3D Object Reconstruction from a single 2D image is an ill-posed question [5], [9], [10], [12], [13], [21] because the in-depth information and distance information is lost during the photo or image capture process. Also, there is no unique solution since the viewpoints for the image can exit on too many possible positions, and the viewpoints can also be changed simultaneously with the distance between the camera and the target object. Furthermore, the 3D Object Reconstruction from a single-view image needs vertical and horizontal distance information for all the objects in the scene. Especially when occlusion and noisy complex cases happen in the image, hallucinating the lost information becomes even more challenging [11].

Although this area of research can be classified in many ways, for instance, in [5], it is classified by 3D shape representation, which is based on whether the 3D representation is Euclidean or non-Euclidean, and based on this, they mainly divide all the research in this area into two groups: Euclidean volumetric approaches and Non-Euclidean/geometric approaches; however, we classify this area's most current state-of-the-art research into two main streams as follows: 1. Based on 3D Ground Truth as the input. 2. Directly infer from 2D single-view image approaches (which means not based on 3D ground truth input). The first kind of research like [12], [13], [14], uses the 3D Ground Truth as the input, and those 3D ground truth includes ready-made CAD 3D models, or the ready-made 3D representations like 3D point cloud, 3D meshes, 3D octrees or 3D voxels, then based on those 3D ground truth input, combined with the 2D image, they inference the 3D object shapes and poses. The second kind of research, like [11], [15], [16], is directly doing 3D reconstruction from the 2D single image; they usually use the semantic segment results from the 2D image and get the RoI (Region of Interest) area which includes the target object, and then use other methods and neuronal networks to first convert it into voxels or meshes and then covert and refine it into 3D triangle meshes, and finally get the shape, poses and other 3D semantics of the scene.

It's important to acknowledge that while the outlined categories provide a structured understanding of the various methods, they are not exhaustive and may not encapsulate all the research contexts or scenarios. Some studies may mixture multiple technologies, and overlaps between categories are common, especially when applied to specific experiments or projects. This classification is not rigid but rather a versatile framework designed to progress in step with ongoing innovations in 3D reconstruction research. In this paper, we will analyse the most recent state-of-the-art papers and research results in both directions in part two and part three;

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00010

and then, the best optimal research methodologies and datasets, as well as a prediction about the future trend, will be suggested in part four; and the conclusion part will summarise all the deliverables for this paper.

II. 3D Ground Truth Model as Input Approaches

Although this group uses different 3D representations in their research, they depend on the indispensable 3D ground truth as the input. The 3D ground truth input representations can be but are not limited to Voxels, Octrees, point clouds, meshes, primitive shapes, and implicit surfaces. All those 3D ground truths are mainly collected by Lidar sensors, 3D depth cameras or manually created using software like CAD (Computer-Aided Design).

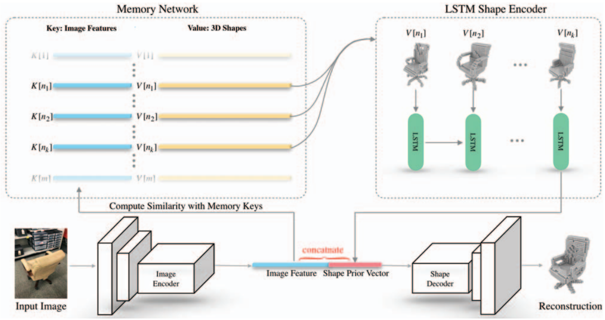

Figure 1. For the Mem3D [17] neural network: the ground truth shapes are stored in a memory network and used later to help recover the occluded parts of the 3D shape.

For the research in [17], they proposed a Mem3D deep learning neural network, as shown in Figure 1, that stores the 3D ground-truth shape priors' information in the form of "image-voxel" pairs into the LSTM architecture memory network, which simulates the human's memory process when they reconstruct the natural world in their brains. For the "image-voxel" pair, a key-value pair, the key refers to the image features collected from the process of 2D CNN in the image encoder stage. The value part refers to the voxel representations of the 3D Shapes of the target object. So, during training, the image features and 3D shapes are stored in the memory network. For both the training and testing stage, the LSTM shape encoder would generate the shape prior vectors for similar 3D shapes, and those shape prior vectors would finally be used in the shape decoder stage to generate the final 3D Reconstruction result. For the decoder part, it uses both the shape prior vectors and the 2D image feature maps to reconstruct the 3D shapes of the target object, more importantly, to recover the occluded parts, the hidden parts, and noisy parts.

By using this Mem3D neural network, they have effectively solved and alleviated the heavy occlusion issues, the noisy environment issues, and the complex scene issues. This approach is inspired and motivated by the human being's vision; humans can reconstruct the 3D scene smoothly for a highly complex scene or heavily occluded objects using their prior knowledge and experiences of the object's 3D shapes, size and poses.

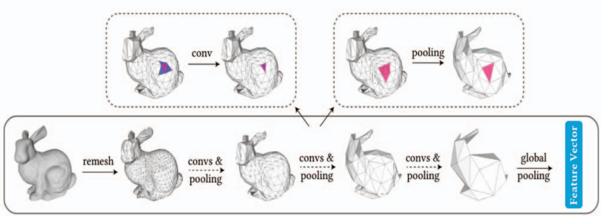

Figure 2. SubdivNet [18]: input a coarse 3D mesh, then do convolutions on vertexes, edges, and surfaces, output a refined mesh

For the research in [18], they proposed a neuronal network named SubdivNet, as illustrated in Figure 2; they firstly imported coarse 3D mesh representations of the 3D object and then did 3D geometric learning on meshes; they used a mesh pyramid structure to do mesh convolution, which includes Vertex-based convolution, Edge-based convolution, and Face-based convolution, which is inspired by computer graphics. This method gives us a good way how to do 3D convolution. Instead of doing a convolution on a 2D plane, 3D convolution usually needs to do convolution on every surface of the cubic object. The way of dealing with 3D convolution and 3D reconstruction is not only flexible, accurate and efficient but also makes sense; when we do computer graphics, we usually draw the vertexes, edges and faces, respectively, and from OpenGL Computer Graphics, we know that every 3D graphics or model can be denoted as a combination of polygons. Those polygons then can be decomposed into a series of triangles. More importantly, they used the Multi-resolution modelling methods to refine the original mesh to make it have subdivision sequence connectivity and refine the original coarse mesh by convolutions.

Method	Accuracy
PointNet++[Qi et al. 2017b]	64.3%
MeshCNN[Hanocka et al. 2019]	92.2%
PD-MeshNet [Milano et al. 2020]	94.4%
MeshWalker [Lahav and Tal 2020]	98.6%
SubdivNet (w/o majority voting)	98.9%
SubdivNet	100.0%

Table 1. SubdivNet [18]: Classification accuracy on the Cube Engraving dataset. SubdivNet is the first method to classify all test meshes correctly.

Although it achieved the state-of-the-art evaluation results on the Cube Engraving dataset depicted and illustrated in Table 1, the SubdivNet neural network is the first to classify all the test meshes correctly. However, this SubdivNet needs a 3D ground truth coarse Mesh as the input instead of inference the 3D shape and poses from a simple 2D image, which leads to SubdivNet's poor generalisation ability. So, when people want to use this method to convert their 2D images to reconstruct 3D objects and sceneries, they need to finish converting their 2D images into a 3D mesh first, and those 3D meshes also need to meet their requirements, and then those 3D meshes can be used as the input, then in the SubdivNet, all those 3D meshes, would be re-meshed, and do all the convolutions and pooling, and then generate the results. Moreover, the flaw also lies in need to ensure the convolution is ordering-invariant before they do the iterate convolution refinement stages, which makes this SubdivNet not very user-friendly. Nevertheless, first thing first, this approach avoided the most challenging part of the 3D reconstruction process, and essentially, the proposed neural network cannot be directly used in inversing the 2D images to 3D shape and pose results, which did not solve the most challenging part of the Single-View 3D Object Reconstruction.

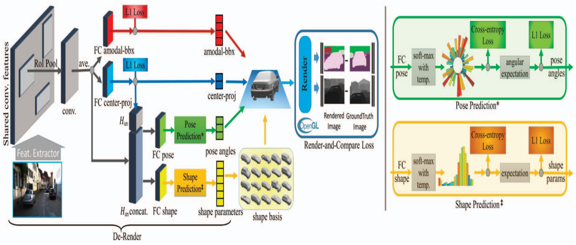

Figure 3. 3D-RCNN [19] network architecture for instance-level 3D object reconstruction

For the research in [19], they made a neural network named: 3D-RCNN, as shown in Figure 3, which uses the ResNet as backbone feature map extraction. In this 3D-RCNN, they proposed a way to do the 2D to 3D inverse graphics by firstly loading in the specified category of 3D objects' ground truth models, which are generated by CAD software; for those 3D CAD models, they have the real-world object's priors' pieces of knowledge. This team exploited the traditional research method of the RoI (Region of Interest) method; since the final pose and shape will be predicted based on the RoI area, which is a segmentation part from the image that includes the target object, they suggest re-parameterise the object from egocentric pose to the allocentric pose, since allocentric orientation is a better representation for learning object orientation as vividly illustrated in Figure 4.

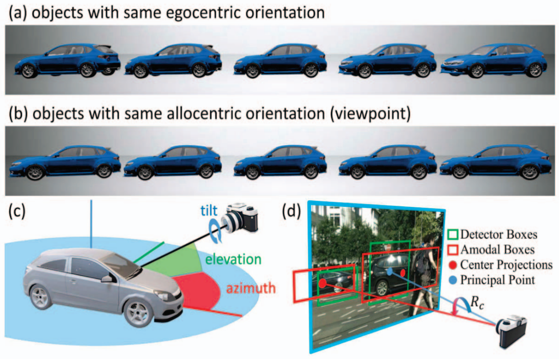
(a) objects with same egocentric orientation
(b) objects with same allocentric orientation (viewpoint)
(c)
tilt
elevation
azimuth
(d)
Detector Boxes
Amodal Boxes
Center Projections
Principal Point
R_c

Figure 4. [19]:(a) the egocentric representation needs to predict the same angle for different image appearances. In (b), all cars in the image have the same allocentric orientation, and we do not see any appearance change. Thus, allocentric orientation is a better representation of learning object orientation. (c) Furthermore, (d) shows the calculation of the camera viewpoint position, the center point of the projection bounding box, the principal point, the amodal bounding box, and the detector box.

Furthermore, they point out that it is essentially and crucially not to ask the network to forecast or calculate the absolute location (Z index or distance) or deep information for the object since it is a fundamentally ill-posed problem; alternatively, instead, they suggest to use their network to estimate the 2D projection of the canonical object centre c = [xc, yc, 1], and the 2D amodal bounding box of the object a = [xa, ya, wa, ha] where (xa, ya) is the centre of the box and (wa, ha) denotes the size of the box. Furthermore, finally, use the network to get a compact 3D parametrisation model of the scene. In this paper, the idea of using the viewpoint concept of computer graphics to analyse, and the idea for the analysis of object relative pose and position part, are both novel and innovative, which is potentially a correct direction for solving the 3D reconstruction issue. However, using the CAD to generate the 3D model prior knowledge part, while impressive, has a fundamental flaw: since it uses the ground truth 3D shape and layout, it relies on 3d solid supervision for training. However, to generate significant, valid, verified, and category-varied training data sets of this kind is impractical, which fundamentally limits the generalisation, scalability, availability and usability of this approach.

III. Directly Infer from 2D Images Approaches (Not Based on 3D Ground Truth Input)

This group of research directly predicted the 3D reconstruction outcome from the single 2D image; in other words, they did not use the 3D ground truth in their research, but alternatively, they used other methods, for example, using Multiview images of the object which from different viewing angles to get the photo of the object and then combine that information to do the 3D reconstruction and estimation inference.

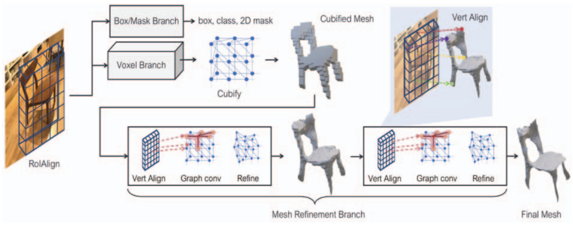

Figure 5. [11]: The overview of Mesh R-CNN.

The research in [11] made a neuronal network named "Mesh R-CNN", which is based on the neuronal network of "Mask R-CNN". They firstly use a voxel branch to predict coarse voxel representations from the region proposal network (RPN) proposals from the Mask R-CNN network, analyse the coarse voxel and convert it into an initial triangle mesh; for the second stage, they use a mesh refinement branch to refine the initial triangle mesh, then output the final 3D mesh representation of the object. Inspired by Mask R-CNN's mask prediction branch, they composed this voxel branch, which predicts a G × G × G grid that gives the object's full 3D shape instead of a 2D M × M bounding box grid of Mask R-CNN.

However, the voxel prediction usually costs too much calculation [11], [20], [21]; even for applying the small fully-convolutional network between the predicted voxels and the input feature map, this fully-convolutional network would take up too much overhead. They use "cubify" tools to convert the coarse voxel into a coarse mesh by analysing the probabilities of voxel occupancy and the binarizing voxel occupancy threshold. After that, they do "graph convolution" on the coarse mesh to refine it and finally get the output 3D object model. During the graph convolution, they only make convolution on vertexes and mesh edges, but they did not do the convolution on mesh triangle faces. As we discussed before, the SubdivNet's convolution method on mesh triangle vertexes, edges, and faces is effective and efficient and made an excellent example of how to do 3D Mesh convolution. If this Georgia's Facebook team could exploit the SubdivNet's 3D Mesh convolution [18], their final refined 3D mesh output would potentially improve.

Furthermore, as is known to all, there exist two ways of object detection: one-stage object detection, like the YOLO (You only look once) [10] series, and the second one is two stages object detection, for example, like Mask R-CNN series. So, for Mask R-CNN, the detection speed is prolonged as a 2-stage object detection method since it needs to generate too many redundant proposals in the first detection stage, retrieve the proposals, and get the best proposal. So, for Mesh R-CNN, built based on the Mask R-CNN, the detection speed would be much slower since it first generates the coarse voxel and uses the small fully-convolutional network to do the mapping. Those operations would consume too much overhead, potentially implying that it is not suitable for applying for 3D real-time scenarios or applications, such as Autonomous driving, Robotics vision. As is known to all, Autonomous driving has an extremely high standard for latency, and only the YOLO series algorithms can almost meet such low latency requirements [10], [15], [18], [21], [22]. However, through human cars-driving behaviours and other daily activities, the speed for human beings of processing 3D object detection and recognition is breakneck and accurate compared to current machine learning, so we infer that there do have a high-speed, efficient and optimal algorithm that can convert the 2D image or scene into 3D model and scene. So, for Mesh R-CNN, it does provide us with thinking of how to convert the 2D image into 3D Scene Estimation; however, it is not the final and optimal solution for 3D Reconstruction and Estimation of a single 2D image.

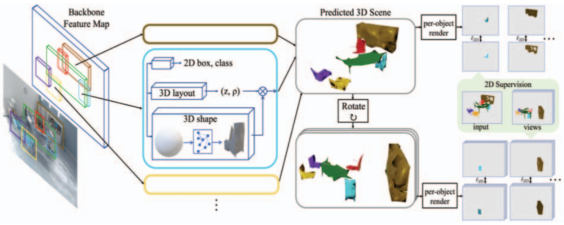

Figure 6. [15]: it takes as input an RGB image, detects all objects in 2D and predicts their 3D location and shape via layout and shape heads, respectively. The output is a scene composed of all detected 3D objects. During training, the scene is differentiable rendered from other views and compared with the 2D ground truth.

For another most recent paper[15], they built the new network based on the previous work of Mesh R-CNN. They firstly pointed out the significant meaning of this area of research, and they stated that it is a fundamental long-last problem in computer vision for the inference of the 3D estimation (including shape and layout) directly from 2D images. Furthermore, the research result has extensive applications as but are not limited to the visions of robotics, autonomous driving, computer graphics, and AR/VR. They admitted that most current research work in this area is only based on the 3D ground truth input of the object, but collecting those 3D ground truth is complex and expensive. Hence, to break the limitations of this dependency, they formed a method of collecting a bunch of 2D images for the specified object from multiple viewpoints using 2D supervision; the proposed neural networks' architect is illustrated in detail in Figure 6. More importantly, their target is to handle and reconstruct many 3D models simultaneously for many individual objects in complex scenes in the realistic world, instead of only dealing with such simple images with only one single object. Consequently, their task is much more challenging and meaningful. The most recent state-of-the-art technology, differentiable rendering, is used in the loss function to learn from multiple views.

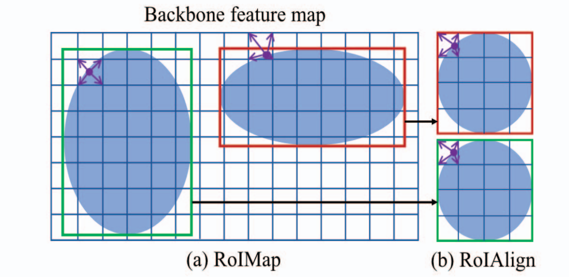

Figure 7. [15] improves the RoIAlign method by using the RoIMap method to extract the features to keep the aspect ratio invariant.

Most excitingly, when computing the vertex-aligned features, they invent a concept of "RoIMap", which can preserve the object's aspect ratio and improve its 3D shape prediction. It is a sagacious and brilliant way to preserve the aspect ratio since, as we all know, in computer graphics, when the images scale up or down, the pixel-wise change can be drastic and dramatic; however, the "aspect ratio" keep unchanged during these scenarios, and it is the most essentially invariant feature when doing the 3D inverse process from a 2D image, and it is also most crucial in all kinds of data augmentation solutions. Although the "RoIMap" approach is very innovative for their research and keeps the aspect ratio of the original object, it does not do the edges and surface convolution for the triangle meshes, and it only provides a vertex alignment and mapping correspondence between the triangle mesh and the original 2D image. Consequently, it caused the designated reconstructed 3D objects to lack sufficient information for a smooth mesh representation, and consequently, their output shapes are of deficient quality. Moreover, as discussed before, this research is based on Mask R-CNN, a two stages object detection algorithm, which implies it would have too big a latency to apply it for scenarios like autonomous driving [22], [23], [24], [25], [26].

IV. Discussion

Remember when humans reconstruct the 3D real world in human brains or when they start to do object detection and recognition; the human cortex mostly learns and processes from their prior 3D objects experiences [12], [13], [14]. Two-dimensional objects take a relatively smaller proportion or rare cases in the real world. Even when humans see the pictures or images when they learn or read or infer from those images, they not only do the pixels wise feature extraction, which is far from sufficient but also, more importantly, the human cortex would reconstruct the 3D representation of the objects and scenes captured in the image by using prior experience, knowledge or imagination. Furthermore, they would try to understand and learn from these images. By analogy, for machine learning or deep learning, when the machines do the object detection task, they should not only extract the pixel-wise feature maps but also have a stage where the machine needs to think about the object's 3D structure and 3D representation from their memory as well. Here, we call it "Machine Imagine" or "3D Scene Semantic Reconstruction" instead of "Machine Learning". So, via this machine imagine, the machine would reconstruct the 3D model for the objects and scenes in the image and then do the inference and mapping of the object in this image to the category in the real world.

Dataset	Image Type	No. of images	No. of categories	Total No. of models
KITTI[27]	Real	12919	11	93,000+
Ikea Dataset[28]	Real	759	7	219
Pascal3D+[29, p. 3]	Real	30,899	12	79
ShapeNetCore	Synthetic (rendered)	-	55	51,300
ModelNet	Synthetic (rendered)	-	662	127,915
ObjectNet3D	Real	90,127	100	44,147
Pix3D[30]	Real	10,069	9	395
ABC	Synthetic (rendered)	-	NA	1,000,000+

Table 2. Standard datasets are used in single-view 3D reconstruction models.

For 2D computer vision, such as object detection and recognition, the most famous and dominant datasets are ImageNet [7] and Microsoft COCO [8]; those two datasets have significant scale and volume of annotated data that accelerate the development and advance of the 2D perception. There is no such dominant dataset for the 3D shape and pose prediction dataset due to the extreme difficulty in collecting 3D annotations [11], [15]. For standard datasets used in single-view 3D reconstruction models are listed in table 1. It seems that the datasets most usually used in the 3D evaluation are: KITTI [27], Pix3D [30], and PASCAL3D+ [29] are the most popular datasets in this area. Other datasets like ShapeNet [31], IKEA [28] dataset, PASCAL3D and, Scene-Shapes, Hypersim [32] and ScanNet [33], and NYUDV2 [34]. For KITTI it has more than 93k out-door street scenes photos which are captured by driving cars; KITTI data have 3D bounding boxes, which have depth maps with a resolution of around 1240×374; however, they do not have annotations for the shapes of the vehicles which makes it very hard to use it for the shape training and prediction. For Pix3D, although it provides a large amount of data, all those objects in the data are indoor objects, mainly IKEA furniture like desks and chairs, which makes it very difficult to use it in other real-world scenarios. For this 3D research area, the lack of highly standardised and high-quality datasets handicaps the research development in this area, all of the datasets in this area have their different standards and benchmarks, which makes it relatively difficult to train and test the data. So, in the future, building a 3D high-quality dataset (like ImageNet or MS COCO in the 2D research area) would be highly recommended and should be the most priority.

From the most recent researches deliverables in this area, it seems there are no most efficient, enjoyable, dominant, and optimal approaches yet for solving the 3D reconstruction and estimation for single-view 2D images; most of the approaches and methods need a 3D ground truth as an input; however, the 3D ground truth fully supervised and annotated data is difficult to collect in large scales, and it is a very challenging task to generate a 3D large scale annotated datasets like the kind of 2D datasets like ImageNet, which made their research no considerable generalisation ability and cannot work in the wild. Some other researchers tried to do the 3D reconstruction and estimation directly without the 3D ground truth; although the idea is innovative and creative, the 3D shapes they generated are too coarse and not very like the original object's silhouettes. Moreover, the model processing time for their method is time-consuming for the step from 2D image segmentation to the coarse 3D voxels representation, which makes it not suitable and feasible for real-time applications like Autonomous driving. However, during the analysis and literature review of the most recent searches, it seems that for the 3D mesh convolution part, the method used in [18] is an efficient way to do the 3D convolution, which involves vertex-based convolution, edges-based convolution and surface-based convolution, which generated a smooth state-of-the-art experiment result of 3D object shape and pose. However, for the 2D Image graphically-reverse to the 3D voxel or mesh stage, it seems that using the computer graphics knowledge and methods to set the viewpoint positions and use the aspect ratio concept to reverse the image becomes the most promising way. The Multiview images approach can be instrumental in getting information from all kinds of viewing angles of the object; however, when fusing those images to generate the 3D models, it needs heavy calculation and is not that efficient. Moreover, inspired by [6], [35], [36], [37], [38] and much other autonomous driving research, we believe that when using the Multiview of 2D images approach, it is recommended that the 2D images firstly map to the Birds-Eye-View (BEV) representations space and then do the 3D reconstruction on the BEV space, since the Birds-Eye-View space vector is much easier to mapping to final 3D Reconstruction vector space, and which would potentially make it easier to reconstruction and estimation the 3D shapes and poses.

From the inspirations of Computer Graphics (like OpenGL principles), there exists an observer for every image when the image is captured; it can be a camera, a person, or a video recorder. Furthermore, it is no arguing that a photograph's formation must have a viewpoint and a distance from the observer to the real object. So, when the camera zooms in or out, the object in the image scales up or down correspondingly. So, when we want to do the 3D reconstruction from a single 2D image, we must consider the viewpoint's position, the object's silhouettes and the distance between the observer and the object. However, currently 2D object detection and recognition, the computer or machine cannot retrieve such features or information from the 2D feature map to detect the same target object when the image size scales up or down. For the machine to have a unified cognition of objects at different scales, it needs to consider the invariant characteristics, like scale invariant and translation invariant characteristics and features of images at different scales.

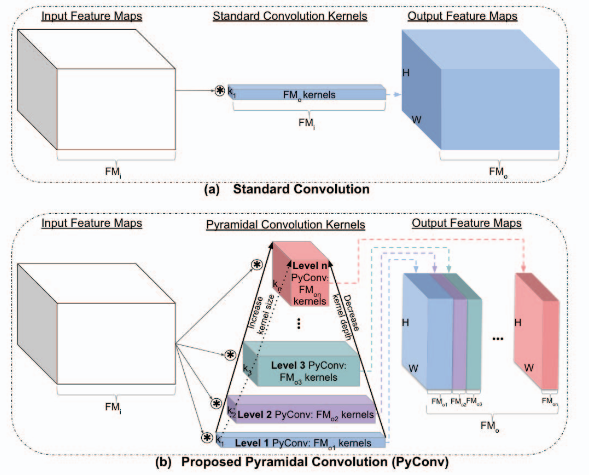
Input Feature Maps
Standard Convolution Kernels
Output Feature Maps
k_1
FM_o kernels
FM_i
H
W
FM_i
FM_o
(a) Standard Convolution
Input Feature Maps
Pyramidal Convolution Kernels
Output Feature Maps
Level n PyConv: FM_{on} kernels
k_n
Increase kernel size
Decrease kernel depth
Level 3 PyConv: FM_{o3} kernels
k_3
Level 2 PyConv: FM_{o2} kernels
k_2
Level 1 PyConv: FM_{o1} kernels
k_1
FM_i
H
W
FM_{o1} FM_{o2} FM_{o3}
FM_{on}
FM_o
(b) Proposed Pyramidal Convolution (PyConv)

Figure 8. [39]: (a) Standard conv; (b) Proposed PyConv.

On this occasion, we believe that The SIFT [40](Scale Invariant Feature Transform) algorithm can help to solve this difficulty. In SIFT algorithm it has a module named "Gaussian Pyramid", which can retrieve different resolutions features of the same image at different scales that simulates the principles of the human visual system because the Gaussian Pyramid has many layers; the bottom layers of the pyramid, it denotes the high resolution and large size and scale of the image, which is mapping to the case of the short distance between the viewpoint and the object; while in the upper layers, it denotes the low resolution and small size of the image, which is mapping to the case of the far distance between the viewpoint and the target object. So, when using Gaussian Pyramid to extract the feature map of the image, it will extract more information than the traditional 2D CNN convolution process, which will give us the depth of information concept of the object, and that will help us a lot in the 3D reconstruction process of the object and the scene. A similar concept has already been used in the pyramidal convolution [39] (PyConv) neural network, so in PyConv neural network, they use a similar concept to Gaussian Pyramid, illustrated in figure 8. When doing the convolution, they made a Pyramid structure that includes several layers; during each layer, it extracts the feature map by different kernel sizes, so in the bottom layers, they use the smallest kernel size, and in the upper layer, they use different increasing kernel sizes. Simultaneously, the kernel depth decreases from the bottom to the top of the pyramid. To do the convolution for different depths of features, the input features are split into the different groups based on the different kernel depths and then do the convolution respectively and correspondingly. For the Standard convolution process, we can see that there is only one feature map extraction process, and it lost too much depth and relative location information, which are necessities for 3D reconstruction. To make accurate 3D Reconstruction results from 2D images, the SIFT algorithm and the Gaussian Pyramid can retrieve much more useful scale invariant and translation invariant information than the current 2D CNN feature map extraction process [39] and consequently would be very promising in generating better results.

V. Conclusion

In this comprehensive survey paper, we analysed the issues in 3D reconstruction and estimation from single 2D images and emphasised the importance of 3D object reconstruction and scene understanding for real-world applications. The information loss during 2D image camera captures and 2D CNN convolutions have also been discussed. All the methods in this research area are divided into two categories: the first category is based on the 3D ground truth input, and the second category does not base on the 3D ground truth. Alternatively, they use other methods to reconstruct the 3D shape directly and pose from the 2D image. Although the first category may generate the state of art results, the requirement of 3D ground truth input has limited their generalisation ability for their approach. The second category would become the future trend for this research field; however, apparently, there is still too much improvement space for it; for instance, the calculation takes too much latency and the experiment's 3D shape and pose results are too coarse, which makes it unrealistic to apply it into the real world and real-time applications, like autonomous driving. Moreover, the standard datasets in this area are discussed, and a large-scale, unified, high quality and highly standardised dataset, among which 3D shape and pose are fully annotated, needs to be built to accelerate the development of this area of research. Last but not the least, the Birds-Eye-View (BEV) representations and SIFT algorithms are discussed and suggested as the methodology and inspirations for the future trend of research in this area.

References

[1] H. Lee, "Convolutional Neural Network," p. 35.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." arXiv, Dec. 10, 2015. doi: 10.48550/arXiv.1512.03385.

[3] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," *ArXiv190505055 Cs*, May 2019, Accessed: Jan. 09, 2022. [Online]. Available: http://arxiv.org/abs/1905.05055

[4] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain Adaptive Faster R-CNN for Object Detection in the Wild," *ArXiv180303243 Cs*, Mar. 2018, Accessed: Jan. 09, 2022. [Online]. Available: http://arxiv.org/abs/1803.03243

[5] G. Fahim, K. Amin, and S. Zarif, "Single-View 3D reconstruction: A Survey of deep learning methods," *Comput. Graph.*, vol. 94, pp. 164–190, Feb. 2021, doi: 10.1016/j.cag.2020.12.004.

[6] Y. Ma *et al.*, "Vision-Centric BEV Perception: A Survey." arXiv, Aug. 04, 2022. doi: 10.48550/arXiv.2208.02797.

[7] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge." arXiv, Jan. 29, 2015. doi: 10.48550/arXiv.1409.0575.

[8] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," *ArXiv14050312 Cs*, Feb. 2015, Accessed: Nov. 03, 2021. [Online]. Available: http://arxiv.org/abs/1405.0312

[9] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D Mesh Renderer," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 3907–3916. doi: 10.1109/CVPR.2018.00411.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *ArXiv150602640 Cs*, May 2016, Accessed: Apr. 18, 2022. [Online]. Available: http://arxiv.org/abs/1506.02640

[11] G. Gkioxari, J. Johnson, and J. Malik, "Mesh R-CNN," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 9784–9794. doi: 10.1109/ICCV.2019.00988.

[12] A.-Q. Cao and R. de Charette, "MonoScene: Monocular 3D Semantic Scene Completion." arXiv, Mar. 29, 2022. doi: 10.48550/arXiv.2112.00726.

[13] A. Grabner, P. M. Roth, and V. Lepetit, "Location Field Descriptors: Single Image 3D Model Retrieval in the Wild." arXiv, Aug. 07, 2019. doi: 10.48550/arXiv.1908.02853.

[14] S. Yang, M. Xu, H. Xie, S. Perry, and J. Xia, "Single-View 3D Object Reconstruction from Shape Priors in Memory." arXiv, Mar. 04, 2021. doi: 10.48550/arXiv.2003.03711.

[15] G. Gkioxari, N. Ravi, and J. Johnson, "Learning 3D Object Shape and Layout without 3D Supervision." arXiv, Jun. 14, 2022. doi: 10.48550/arXiv.2206.07028.

[16] Z. Li *et al.*, "Neuralangelo: High-Fidelity Neural Surface Reconstruction." arXiv, Jun. 12, 2023. doi: 10.48550/arXiv.2306.03092.

[17] S. Yang, M. Xu, H. Xie, S. Perry, and J. Xia, "Single-View 3D Object Reconstruction from Shape Priors in Memory," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 3151–3160. doi: 10.1109/CVPR46437.2021.00317.

[18] S.-M. Hu *et al.*, "Subdivision-Based Mesh Convolution Networks," Jun. 2021, doi: 10.1145/3506694.

[19] A. Kundu, Y. Li, and J. M. Rehg, "3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 3559–3568. doi: 10.1109/CVPR.2018.00375.

[20] J. S. Murthy, G. M. Siddesh, W.-C. Lai, B. D. Parameshachari, S. N. Patil, and K. L. Hemalatha, "ObjectDetect: A Real-Time Object Detection Framework for Advanced Driver Assistant Systems Using YOLOv5," *Wirel. Commun. Mob. Comput.*, vol. 2022, p. e9444360, Jun. 2022, doi: 10.1155/2022/9444360.

[21] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why Reinventing a Face Detector," *ArXiv210512931 Cs*, May 2021, Accessed: Oct. 17, 2021. [Online]. Available: http://arxiv.org/abs/2105.12931

[22] C. Liu, H. Sui, J. Wang, Z. Ni, and L. Ge, "Real-Time Ground-Level Building Damage Detection Based on Lightweight and Accurate YOLOv5 Using Terrestrial Images," *Remote Sens.*, vol. 14, no. 12, Art. no. 12, Jan. 2022, doi: 10.3390/rs14122763.

[23] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," arXiv, arXiv:2107.08430, Aug. 2021. doi: 10.48550/arXiv.2107.08430.

[24] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." arXiv, Jul. 06, 2022. Accessed: Jul. 14, 2022. [Online]. Available: http://arxiv.org/abs/2207.02696

[25] A. Benjumea, I. Teeti, F. Cuzzolin, and A. Bradley, "YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles." arXiv, Dec. 23, 2021. doi: 10.48550/arXiv.2112.11798.

[26] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios," *ArXiv210811539 Cs*, Aug. 2021, Accessed: Jan. 09, 2022. [Online]. Available: http://arxiv.org/abs/2108.11539

[27] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013, doi: 10.1177/0278364913491297.

[28] Y. Ben-Shabat *et al.*, "The IKEA ASM Dataset: Understanding People Assembling Furniture through Actions, Objects and Pose." arXiv, Jul. 01, 2020. doi: 10.48550/arXiv.2007.00394.

[29] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A benchmark for 3D object detection in the wild," in *IEEE Winter Conference on Applications of Computer Vision*, Mar. 2014, pp. 75–82. doi: 10.1109/WACV.2014.6836101.

[30] X. Sun *et al.*, "Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling." arXiv, Apr. 12, 2018. doi: 10.48550/arXiv.1804.04610.

[31] A. X. Chang *et al.*, "ShapeNet: An Information-Rich 3D Model Repository." arXiv, Dec. 09, 2015. doi: 10.48550/arXiv.1512.03012.

[32] M. Roberts *et al.*, "Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding." arXiv, Aug. 17, 2021. doi: 10.48550/arXiv.2011.02523.

[33] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes." arXiv, Apr. 11, 2017. doi: 10.48550/arXiv.1702.04405.

[34] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *Computer Vision – ECCV 2012*, vol. 7576, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., in Lecture Notes in Computer Science, vol. 7576. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 746–760. doi: 10.1007/978-3-642-33715-4_54.

[35] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View." arXiv, Jun. 16, 2022. Accessed: Aug. 13, 2022. [Online]. Available: http://arxiv.org/abs/2112.11790

[36] Y. Zhang *et al.*, "BEVerse: Unified Perception and Prediction in Birds-Eye-View for Vision-Centric Autonomous Driving." arXiv, May 19, 2022. doi: 10.48550/arXiv.2205.09743.

[37] E. Xie *et al.*, "M^{2}BEV: Multi-Camera Joint 3D Detection and Segmentation with Unified Birds-Eye View Representation." arXiv, Apr. 19, 2022. doi: 10.48550/arXiv.2204.05088.

[38] Y. Zhao, Y. Zhang, Z. Gong, and H. Zhu, "Scene Representation in Bird's-Eye View From Surrounding Cameras With Transformers," p. 9.

[39] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Pyramidal Convolution: Rethinking Convolutional Neural Networks for Visual Recognition." arXiv, Jun. 20, 2020. doi: 10.48550/arXiv.2006.11538.

[40] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, Greece: IEEE, 1999, pp. 1150–1157 vol.2. doi: 10.1109/ICCV.1999.790410.