State of the Art on: 3D object reconstruction

Cristian Sbrolli, cristian.sbrolli@mail.polimi.it

1. Introduction to the research topic

During the last decade, deep learning models arose and reached remarkable state-of-the-art results in many machine learning tasks, gaining a central spot in the research field. As analyzed in [2], deep learning has many successful applications in the real world, including natural language processing, medical techniques and computer vision. Deep learning models leverage deep neural networks to learn to extract useful features from data, which are then used to perform a specific task. A relevant topic in the field of computer vision is 3D reconstruction, which goal is to infer the 3D geometry and structure of objects (or even scenes) from inputs ranging from one or multiple 2D images to 2.5D images to point clouds. In particular, single view 3D reconstruction poses several challenges due to the loss of information in the 3D to 2D projection. This long standing ill-posed problem is fundamental to many applications such as robot navigation, object recognition and scene understanding, 3D modeling and animation, industrial control, and medical diagnosis. Humans are good at solving such ill-posed inverse problems by leveraging prior knowledge. Indeed, we can easily infer the approximate size and rough geometry of objects and even guess what it would look like from another viewpoint, all of this just by looking at a single image. We can do this because all the previously seen objects and scenes have enabled us to build prior knowledge and develop mental models of what objects look like. Before deep learning, this problem has been approached by a geometric prospective, focusing on understanding and formalizing, mathematically, the 3D to 2D projection process, with the aim to devise mathematical or algorithmic solutions to the ill-posed inverse problem. With deep learning instead, the approach shifted to that of trying to leverage the previously described human ability to learn prior knowledge. Despite having demonstrated promising results, one of the relevant problems of the earlier deep learning architectures for 3D reconstruction is the need of annotated 3D targets for the supervised training of these networks, which are harder to collect w.r.t. 2D annotations. As a solution, some works started shifting the target of the task from 3D to 2D by incorporating into the network architectures differentiable renderers (1.1.2), which allow to maintain an end-to-end training of the architecture, now including a rendering process from a 3D model to a 2D one, enabling the use of 2D annotations, which are easier to obtain. This allowed the birth of new models leveraging self-supervised learning, which has been an emerging research topic in the recent years. Other main problems afflicting 3D reconstruction models from single images are occlusion and noisy backgrounds: occlusion affects the quality of reconstruction of unseen parts of the input image, while noisy backgrounds can in general prevent the reconstruction of the object or some of its parts if the background mixes with them or causes an unclear visual segmentation of the object. As mentioned above, both problems are easily solved by humans by leveraging prior shapes, so recent works tried to learn explicit shape priors or leverage prior shapes from databases so to emulate this human ability, obtaining relevant improvements w.r.t. the state of the art models. To stress the importance of the topic and the issues described above, we also include a list of some of the most reputable publication journals, associated to the field:

Conference Name	**Impact Factor**
CVPR: Conference on Computer Vision and Pattern Recognition	51.98
ECCV: European Conference on Computer Vision	25.91
ICCV: International Conference on Computer Vision	32.51
BMVC: British Machine Vision Conference	5.94
3DV 2021 : International Conference on 3D Vision	5.94
SIGGRAPH: Eurographics Symposium on Computer Animation	1.98

1.1. Preliminaries

1.1.1 Space representations

When dealing with 3D shapes, different representations with different characteristics can be used.
Voxels are the natural 3D extension of 2D pixels. Each voxel, identified by discrete coordinates xyz, represent a value on a regular grid in three-dimensional space, and can hold a set of features as color, material and transparency. Voxel representations are dense, thus less memory efficient, but allow easier rendering and transformations.
Polygon meshes represent 3D shapes as a set of vertices and the surfaces connecting them, usually triangular faces. It is a sparse representation, since only the surfaces are represented, while empty regions and the internal volume are not represented.
Point clouds are collections of points in 3D space; each point is specified by an xyz location, along with some attributes (e.g., color). They are also sparse representations, since only non-empty regions are represented.
Implicit representations. Some recent models, as [7], started representing 3D information in a parametrized manner in neural networks, commonly known as neural implicit representation. In this model, the geometric information at a point $p \in \mathbb{R}^3$ is described by the output of a neural network $f(p)$, which takes the latent representation of a shape and a 3D point, and returns a value indicating whether the point is outside or inside the shape.

1.1.2 Differentiable rendering

Rendering, in computer graphics, is the process of generating images of 3D scenes defined by geometry, materials, scene lights and camera properties. Rendering is a complex process and its differentiation is not uniquely defined, which prevents straightforward integration into neural networks. Differentiable rendering (DR) constitutes a family of techniques that tackle such integration, for end-to-end optimization, by obtaining useful gradients of the rendering process. By differentiating the rendering, DR bridges the gap between 2D and 3D processing methods, allowing neural networks to optimize 3D entities, while operating on 2D projections. Instead of handcrafting differentiable renderers, [3] propose *neural rendering*, which is learning the rendering process from data. This is generally realized by jointly training the shape reconstruction network and the rendering network on the image reconstruction error. By learning on real world data, neural rendering can produce images which are indistinguishable from real images, on the other hand handcrafted renderers images are usually less accurate; however, neural rendering suffers from some problems, among which the one of losing accuracy when generalizing to images differing from training data. Some works, as [9], highlight that a promising direction for improving neural rendering is adding inductive biases for 3D scenes, on which handcrafted rendering are based, thus combining the two approaches.

1.2. Research topic

The aim of the proposed work is that of exploring and enhancing the ability of 3D reconstruction models to leverage and learn 3D shape priors, while exploiting recent architectures as graph networks and transformers. The former have been proved to be applicable for exploiting prior shapes as in [11], while the application of the latter to the 3D reconstruction problem is still unclear and not well explored. Indeed, as shown in [14], leveraging explicit priors by a sort of attention mechanism, significantly outperforms other state-of-the-art models, which suggests that a transformer approach may be effective. In [6], state-of-the-art performances are achieved with the aid of an explicit prior retrieved from a database, then re-modelled by free-form deformation. All these approaches present interesting and effective tools to face the challenges of 3D reconstruction, leading to the question this work aims at finding an answer to: if (and how) can these approaches be combined and enhanced to effectively leverage prior shapes to solve the problems of single view reconstruction? We can also pose another interesting question: can something more be inferred by exploiting parametrized database models? As an example, given an image of an object and a prior shape (3D model) of an object of the same category, can we infer the parameters to apply

to the 3D model to obtain the object in the image? Answering these questions, especially the last one, would be useful in many applications as in procedural generation of scenes.

2. Main related works

2.1. Classification of the main related works

Models found in literature can be classified by the nature of their input, by the representation of their output and by their architecture. Though we focus on single input images, some models can handle multiple images, video frames or even 3D representations as point clouds and voxels. Output representation is crucial to aspects such as efficiency and reconstruction quality.
Volumetric representations represent space by regular voxel grids, which allow 3D convolutions and so an easier extension from 2D techniques at the price of memory efficiency. *Surface-based representations*, which includes representations as point clouds and meshes are more memory efficient, but cannot fit into classic deep learning architectures and require to be treated specifically. Another relevant aspect is the presence of a prior for the shapes and, if present: (1) which kind of prior is used, (2) if it is learned or not, (3) if it iS general or category-specific, (4) and whether it is explicit or implicit. Other factors to classify systems for neural-based reconstruction are training procedure, grade of supervision and the used datasets (synthetic or wild).
The most used network architecture is the encoder-decoder architecture, in which the image is first embedded in a latent space by the encoder, then reprojected (decoded) into the desired output dimension.
Among the state-of-the-art models with no shape prior, one can find the following works.:
[4] uses single image input, a point cloud representation and training is achieved by predicting 2D views of the image with a weakly supervised framework.
[13] can use a single or multiple images as input, a voxel representation and supervised 3D training.
[7] and [1] can take a single image as well as point clouds or voxels as input, learn an occupancy function assigning occupancy probability to points, which is then used to extract 3D meshes and trains with 3D supervision.
[15] can use single or multiple images as input, leverages a transformer network to predict voxel representations and supervised 3D training.
Among the models trying to exploit a prior shape:
[12] uses a single input image, a voxel representation and it exploits adversarial learning implicit shape priors.
[5] uses a single input image, a mesh representation and it trains by adversarial learning shape priors on projected 2D shapes.
[14] uses a single input image, voxel representation, 3D supervision and it learns explicit shape priors.
[6] uses a single input image, point cloud representation, 3D supervision and explicit database shape priors.

2.2. Brief description of the main related works

Weakly-Supervised Single-view Dense 3D Point Cloud Reconstruction via Diferentiable Renderer [4]
The model presented takes a *single RGB image* as input and learns to reconstruct the 3D point cloud representation of the object in the image. An encoder-decoder architecture is leveraged to encode the image in a latent space; the hidden representation is then used to predict multiple views of the object, then fused into a single 3D point cloud. A differentiable renderer is then used to predict novel views from the reconstructed model (2.5D images) which, compared to the ground truth images for the respective views, allow to compute a loss. The system also comprises a pose estimation CNN used for estimating the poses of the ground truth images, which are used by the differentiable renderer to predict the views. The model is trained first on a synthetic dataset, then fine tuned on a real world dataset. The main drawbacks of this approach are the need of 3D annotations and the fact that no prior shape is considered to enhance the results..

Pix2Vox: Context-aware 3D Reconstruction from Single and Multi-view Images [13]
The proposed model works as follows: each input image, in parallel, pass through an encoder-decoder architecture, generating a coarse 3D volume. The resulting 3D volumes are forwarded to the multi-scale context-aware fusion module that adaptively selects high-quality reconstructions for different parts from all coarse 3D volumes in parallel, generating a fused 3D volume. Finally, a refiner further corrects the wrongly recovered parts of the fused 3D volume to produce the final reconstruction. The training is supervised and it uses voxel-wise binary cross entropies between the reconstructed object and the ground truth. The main limitations of this approach are the same as for the first work described above..

OccNet and D-OccNet [7] and [1]
The Occupancy Network is an encoder-decoder convolution neural network architecture. The input is encoded into a vector and the decoder learns an occupancy function which takes the image embedding and N 3D points, assigning an occupancy probability to each input point. This allows the model to extract meshes at any arbitrary resolution, using a multi-resolution isosurface extraction algorithm. D-OccNet extends this work by using two pipelined OccNets. The first one takes an image and it returns a mesh, which is then transformed into a point cloud given as input to the second OccNet, wich returns the final mesh. The standard (single) OccNet can only leverage the image 2D information, while the advantage of D-OccNet is that of allowing the second OccNet to also exploit the 3D information contained in the generated point cloud, which leads to a much finer quality in the reconstruction, though the quantitative results are similar due to the evaluation metrics not considering fine details. Training needs 3D supervision and it involves sampling the 3D bounding volume to evaluate the binary cross entropy of points.

3D-RETR: End-to-End Single and Multi-View 3D Reconstruction with Transformers [15]
This is the first work using a transformer model to perform 3D end-to-end reconstruction and reach state-of-the-art performances. The presented architecture is composed by a pretrained transformer to extract visual features from 2D images, a transformer decoder to extract voxel features from visual features, a CNN decoder producing the 3D representation from the voxel features. The model trains with supervision by 3D ground truth voxel models, but the classic binary cross entropy loss is replaced with Dice loss, which helps when the metric to optimize is IoU (Intersection over Union).

Learning Shape Priors for Single-View 3D Completion and Reconstruction [12]
The presented model (ShapeHD) aims at learning an implicit shape prior in the form of a *Naturalness score*, which penalizes the model if its output is unrealistic, which can happen in occluded or noisy parts.This is achieved by an adversarial approach, where the discriminator from a pre-trained 3D generative adversarial network is used to determine whether a shape is realistic, thus modelling the afore mentioned *Naturalness score* by having the ability to model the real shape distribution. The model uses two encoder-decoder networks: one to generate 2.5D sketches from a 2D image and one taking the 2.5D image sketches and producing the 3D shape (as voxel); the generated shape is then fed to the discriminator. The training makes use of a combined loss including both a loss on the 3D shape (voxel-wise binary cross entropy) and a loss on naturalness of the shape.

Learning View Priors for Single-view 3D Reconstruction [5]
This work, similarly to the just discussed ShapeHD, exploits an adversarial network to learn implicit priors. The difference lies in the fact that ShapeHD learns priors over 3D models by the *Naturalness score*, while this model learns priors over 2D views. This is achieved by training a discriminator that learns prior knowledge regarding possible views: it is trained to distinguish the reconstructed views of the observed viewpoints from those of novel (unobserved) viewpoints. From an architectural point of view, the model uses an encoder-decoder to generate a 3D mesh and a texture, used to render 2D views among which the original view from the input image; the generated views are then fed to the discriminator which has to identify if a view is the original or a novel one. The 3D model is generated by moving the vertices of a pre-defined mesh (e.g. ellipsoidal mesh), therefore the output of the shape decoder are the coordinates of the estimated

vertices. The training involves three loss terms: the discriminator loss, an internal pressure loss on the 3D model and a reconstruction loss on the 2D image rendered from the input view.

Single-View 3D Object Reconstruction from Shape Priors in Memory [14]
The proposed model (Mem3D) explicitly constructs shape priors to supplement the missing information in the image. Specifically, the shape priors are in the forms of "image features-voxel" pairs in a memory network, which is stored by an ad hoc writing strategy during training. The architecture uses an encoder to extract image features, which are then used to retrieve similar shapes from the memory network with an attention-like mechanism using keys and values; the retrieved voxels are then processed by an LSTM network extracting a shape prior vector, which is then concatenated to the extracted features and fed into a decoder generating the 3D reconstruction. The training uses a voxel triplet loss that helps to retrieve precise values from the memory network by guaranteeing that images with similar 3D shapes are closer in the feature space and a binary cross entropy loss on the reconstructed voxel with respect to the ground truth 3D model.

DeformNet: Free-Form Deformation Network for 3D Shape Reconstruction from a Single Image [6]
The model presented in this work uses a single image input to first perform shape retrieval from an object dataset using an image-to-shape embedding learnt by metric learning. Then, it deforms the point cloud representation of the retrieved template using the **FFD** (Free Form Deformation) layer in an encoder-decoder style network architecture. FFD [8] is the 3D extension of a Bezier curve form, which has been widely used for shape deformation; it defines control points over a grid on the 3D shape, allowing to deform the shape by moving (through offsets) the control points. The model is trained in a supervised manner by 3D ground truth models.

2.3. Discussion

In light of the extensive research undertaken in the past years, single view 3D reconstruction has achieved impressive results. The topic, however, still has lots of potential branches to be explored and existing ones to be further analyzed. In particular, we list some of the issues still present:

Training data. Deep learning models heavily depend on the amount of training data. The size and the number of datasets for single view 3D reconstruction which are annotated with the corresponding 3D model are small, compared for datasets for tasks as classification or recognition. As discussed, this problem has been addressed by shifting the problem to 2D supervision, but still, these models mostly use silhouette-based losses, thus they can only reconstruct the visual hull.

Generalization to unseen objects. For most of the state-of-the-art models, it is not clear how they would perform on unseen object/image categories, which instead is a fundamental point on general 3D reconstruction.

In-the-wild-images. Although obtaining impressive results in synthetic datasets, most models heavily lose accuracy when dealing with real world images, due to heavy occlusion or noisy backgrounds. Furthermore, while in synthetic datasets there is a single object per image, real images could contain multiple objects, which are not handled by most models.

Fine-scale 3D reconstruction Although recent models improved the quality of the reconstructed models, they struggle to reconstruct small or thin parts.

Reconstruction vs Recognition As shown in [10], datasets as ShapeNet suffer from a training-test split problem, for which for a typical shape in the test set, there is a very similar shape in the training set. They show this by building state-of-the-art models by only retrieving a similar shape from the training set. Their analysis indicates that state-of-the-art approaches to single-view 3D reconstruction primarily perform recognition rather than reconstruction, which will more likely generalize badly to unseen data.

References

[1] Ansari, M. U., Bilal, T., and Akhter, N. D-occnet: Detailed 3d reconstruction using cross-domain learning, 2021.

[2] Dong, S., Wang, P., and Abbas, K. A survey on deep learning and its applications. *Computer Science Review 40* (2021), 100379.

[3] Eslami, S. M. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., Reichert, D. P., Buesing, L., Weber, T., Vinyals, O., Rosenbaum, D., Rabinowitz, N., King, H., Hillier, C., Botvinick, M., Wierstra, D., Kavukcuoglu, K., and Hassabis, D. Neural scene representation and rendering. *Science 360*, 6394 (2018), 1204–1210.

[4] Jin, P., Liu, S., Liu, J., Huang, H., Yang, L., Weinmann, M., and Klein, R. Weakly-supervised single-view dense 3d point cloud reconstruction via differentiable renderer. *Chinese Journal of Mechanical Engineering 34*, 1 (Sep 2021), 93.

[5] Kato, H., and Harada, T. Learning view priors for single-view 3d reconstruction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 9770–9779.

[6] Kurenkov, A., Ji, J., Garg, A., Mehta, V., Gwak, J., Choy, C., and Savarese, S. Deformnet: Free-form deformation network for 3d shape reconstruction from a single image, 2017.

[7] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[8] Sederberg, T. W., and Parry, S. R. Free-form deformation of solid geometric models. *SIGGRAPH Comput. Graph. 20*, 4 (aug 1986), 151–160.

[9] Sitzmann, V., Zollhoefer, M., and Wetzstein, G. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems* (2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc.

[10] Tatarchenko, M., Richter, S. R., Ranftl, R., Li, Z., Koltun, V., and Brox, T. What do single-view 3d reconstruction networks learn? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 3400–3409.

[11] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., and Jiang, Y.-G. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)* (September 2018).

[12] Wu, J., Zhang, C., Zhang, X., Zhang, Z., Freeman, W. T., and Tenenbaum, J. B. Learning 3D Shape Priors for Shape Completion and Reconstruction. In *European Conference on Computer Vision (ECCV)* (2018).

[13] Xie, H., Yao, H., Zhang, S., Zhou, S., and Sun, W. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision 128*, 12 (Jul 2020), 2919–2935.

[14] Yang, S., Xu, M., Xie, H., Perry, S., and Xia, J. Single-view 3d object reconstruction from shape priors in memory. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 3151–3160.

[15] Zai Shi, Zhao Meng, Y. X. Y. M. R. W. 3d-retr: End-to-end single and multi-view3d reconstruction with transformers. In *BMVC* (2021).