Research Project Proposal: 3D object reconstruction from shape priors

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

1. Introduction to the problem

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. Among the most successful applications of deep learning, impressive results have been achieved for 2D computer vision tasks, leveraging convolutional networks for extracting features. The next step has been that of considering 3D tasks, which pose different challenges as the extension of 2D models to 3D is often not trivial. Among the 3D tasks we find 3D reconstruction, the goal of which is that of inferring the 3D geometry and structure of objects from inputs ranging from one or multiple 2D images to 2.5D images to point clouds. In particular, 3D reconstruction from single image task poses fundamental problems due to the loss of information caused from the 3D to (single) 2D image 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. Two main approaches have been developed to try to integrate this prior knowlege into the models: implicit and explicit shape prior. Implicit prior models aims at learning the concept of "naturalness" of a shape, allowing the model to reconstruct shapes that look realistic even from unobserved viewpoints, so forcing the model to implicitly learn some prior knowledge for each class of object; one way this has been done is by adversarial learning. Explicit prior models instead leverage actual 3D shapes as memory networks storing voxels, biases representing prior shapes or prior models from databases; this explicit priors, combined with the input image features, are then manipulated by techniques as deformation and combination to build the predicted 3D shape. Both approaches are able to effectively leverage the prior knowledge on shapes and to effectively improve the performance w.r.t. models not leveraging prior shapes. Given these progresses by using prior shapes and the introduction of new architectures as graph networks and transformers, we wonder if and how we can combine and enhance this tools, architectures and concepts to move a step further in single view 3D reconstruction. Moreover we pose another inherent and relevant 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

Recent works highlighted that learning or leveraging prior shapes enhances the performance in 3D reconstruction, both using implicit or explicit prior shapes. In [7] and [2], adversarial training is used to learn a *Naturalness* score (which can be seen as learning implicit prior shapes) of the predicted 3D shape, so the loss function is not only penalizing models differing from the true shape but also the ones looking unnatural. In [8], instead, explicit priors are leveraged by building a readable and writable memory network storing voxel shapes, jointly used with an encoder-decoder network to exploit the memory prior shapes with an attention-like mechanism. In [3] and [6], prior (explicit) shapes from a database are retrieved, based on the embedding of the input image, and then modified by Free Form Deformation (FFD) [4], which is the 3D extension of a Bezier curve form, which has been widely used for shape deformation. FFD defines control points over a grid on the 3D shape which allow to deform it by moving the control points through offsets. In particular, [6] represents 3D meshes in a graph-based convolutional neural network and produces correct geometry by progressively deforming an ellipsoid (general prior), considering perceptual features extracted from the input image representations. By using a general prior, the system obtains impressive results, suggesting that the approach of progressively refining a prior model is effective and thus we may be able to leverage it. [9] is the first transformer architecture applied to 3D reconstruction: it uses a transformer encoder, a transformer decoder and a CNN decoder to generate voxel shapes, outperforming state-of-the-art models and showing that transformers can be the next big step in the field. The model in [9] does not make use of any prior shape, which strengthen our interest on combining the two methodologies. As final observations, here are some problems and intuitions which we may want to address in the proposed research: 3D annotations are a limitation of most of the mentioned models, which can be faced by using differentiable rendering/neural rendering, allowing us to train with 2D supervision; predicting (or deforming) meshes seems to bring more effective reconstructions w.r.t. using voxels; more recent architectures, as graph neural networks transformers, are effective but still quite unexplored.

3. Research plan

The proposed research aims at 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 [6], while the application of the latter to the 3D reconstruction problem is still unclear and not well explored. This goal poses some challenges and choices, among which (1) Defining the model: its general architecture and the specific design of its parts (2) How to represent priors: implicitly as in adversarial models, explicitly by either embedding them or using them directly (3) How to use the priors: as kind of features concatenated to the input embedding, deforming them by [4], combining them by learnt parameters or even just predicting some specific pre-defined parameters to deform a prior shape from a parametrized catalogue of objects (4) The training paradigm: supervised or self supervised by using a differentiable rendering? (5) The dataset to use: synthetic, in-the-wild or both? Moreover, we also wonder whether it is possible to reconstruct a full scene by building it objectwise, especially in the case in which we use a catalogue of parametrized prior shapes: the idea is that of reconstructing each object by leveraging shape priors, eventually merging the reconstructed objects to build a large scale scene. This would pose additional challenges, as this is a slightly different task: we do not want to reconstruct exactly the input objects, but "replicate" them by parametrized catalog objects. This requires specific variations on the reconstruction model to predict object parameters and some different evaluation method w.r.t. the standard evaluation on datasets like ShapeNet [1] or Pix3D[5], which encourage the exact reconstruction of objects. The proposed work develops both a more theoretic aspect, by digging into the engineering of effective models, and an applicative aspect, by verifying their results on synthetic and possibly real data. The planned work comprises:

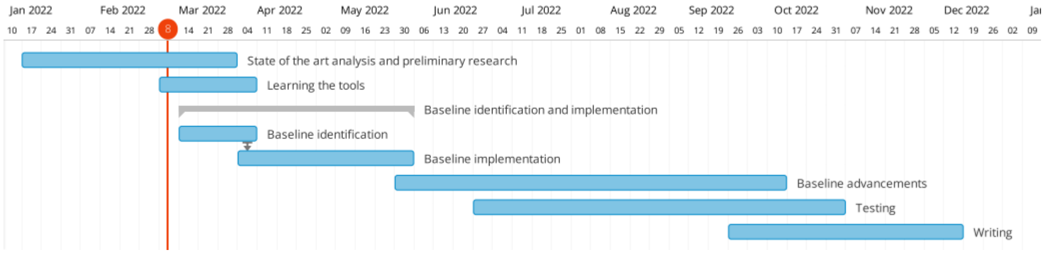
1. *State of the art analysis and preliminary research*: study the literature on 3D reconstruction, particularly on methods leveraging prior shapes. In this phase we identify and classify the possible effective approaches, identifying their strengths and weaknesses.

2. *Learning the tools*: experiment with the existing methods and tools to learn best practises and get used with them.

3. *Baseline identification and implementation*: examine the outputs of the preliminary research and the project goal

to identify the desired approach, implementing a baseline using the chosen approach. This phase includes the setup of the needed tools as the differentiable renderer (if used) and the choice or creation of the dataset.

4. *Baseline advancements*: enhance the baseline model based on its results on the reconstruction task, possibly reconsidering its architecture and features. Consider the extension of the model to a larger scale task and possibly to the mentioned goal of objectwise scene parsing.
5. *Testing*: test the advanced model on public datasets to compare it to current state-of-the-art methods.
6. *Writing*: write the M.Sc. thesis and the conference paper.



The evaluation of the final model will be based on the metrics described in section 3.1, plus possible modified metrics for the large scale and parametrized reconstruction case.

3.1. Evaluation metrics

When evaluating and comparing 3D reconstruction models, we consider firstly quantitative metrics, among which:

The Mean Squared Error (MSE) is defined as the symmetric surface distance between the reconstructed shape X' and the ground-truth shape X: $d(X',X) = \frac{1}{n_X}\sum_{p\in X} d(p,X') + \frac{1}{n'_X}\sum_{p'\in X} d(p',X)$, where n_X and n'_X are, respectively, the number of densely sampled points on X and X', and $d(p,X)$ is the distance, e.g., the L1 or L2 distance, of p to X along the normal direction to X. The smaller this measure is, the better is the reconstruction.

Intersection over Union (IoU) measures the ratio of the intersection between the volume of the predicted shape and the volume of the ground-truth, to the union of the two volumes $IoU(X',X) = \frac{X\cap X'}{X\cup X'}$, which for probabilistic models can be rewritten as $IoU(X',X) = \frac{\sum_i I(X_i>\epsilon)*I(V_i)}{\sum_i I(I(X_i>\epsilon)+*I(V_i))}$ where I(·) is the indicator function and ϵ is a threshold.

Earth Mover's Distance (EMD) and Distance (CD) are designed for point clouds. The EMD is defined as the minimum of the sum of distances between a point in one set and a point in another set over all possible permutations of the correspondences: $d_{EMD}(S_1,S_2) = \min_{\phi:S_1\longrightarrow S_2} \sum_{x\in S_1} ||x-\phi(x)||_2$. In the distance instead, for each point of each point cloud, the nearest neighbor in the other set is found their squared distances is summed up: $d_{CD}(S_1,S_2) = \sum_{x\in S_1}\min_{y\in S_2}||x-y||_2^2 + \sum_{x\in S_2}\min_{y\in S_1}||x-y||_2^2$

Qualitative analysis should not be neglected when evaluating a 3D reconstruction model: the mentioned quantitative metrics (and also the other existing ones) struggle to differentiate the little details between two meshes, which may result in two models having the same quantitative performance, but significantly different ability to capture small details.

References

[1] Chang, A. X., Funkhouser, T. A., Guibas, L. J., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. Shapenet: An information-rich 3d model repository. *CoRR abs/1512.03012* (2015).

[2] 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.

[3] 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.

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

[5] Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J. B., and Freeman, W. T. Pix3d: Dataset and methods for single-image 3d shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[6] 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).

[7] 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).

[8] 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.

[9] 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).