In recent years, Machine Learning techniques have revolutionized solutions to longstanding image-based problems, like image classification, generation, semantic segmentation, object detection and many others. However, if we want to be able to build agents that can successfully interact with the real world, those techniques need to be capable of reasoning about the world as it truly is: a tridimensional space. There are two main challenges while handling 3D information in machine learning models. First, 3D data is not available in the same scale as images – taking pictures is a common procedure in our daily lives, whereas capturing 3D content is an activity usually restricted to specialized professionals. Second, it is not clear what is the best 3D representation for machine learning models. For images, convolutional neural networks (CNNs) operating on raster images yield the best results in virtually all image-based benchmarks. For 3D data, the best combination of model and representation is still an open question. My research is focused in addressing both of these issues. Which model and representation should we use for generating [2, 4] and recognizing [5, 4] 3D data? Is it possible to leverage image data to build models capable of reasoning about the world in 3D [8, 1, 7]?

Our research findings show that we are able to build models that efficiently generate 3D shapes as unstructured point clouds. Those models require significantly less memory while generating higher quality shapes than the ones based on voxels and multi-view representations. These architectures can be applied to shape classification, segmentation, single-view reconstruction and unsupervised representation learning with variational auto-encoders (VAEs). Notwithstanding, the aforementioned techniques require explicit 3D supervision, which is scarce. Ideally, we want to be able to use images as supervisory signal while learning to generate 3D data. We tackle this problem by proposing differentiable neural network modules capable of generating images from 3D representations. The experiments demonstrate that those modules can serve as building blocks to neural network architectures, enabling learning 3D representations from 2D images in a variety of scenarios.

Below, I address the specific topics of my research in more detail.

Learning from Unstructured Data

Tridimensional occupancy grids are a natural choice for representing 3D data in deep neural networks. They are a straightforward extension to raster images and convolutional layers can be seamlessly applied to this type of data. Another way to represent 3D data is by simply utilizing multiple 2D images of a 3D object. We refer to this as multi-view representation. This type of representation can also be easily integrated with convolutional layers and even offers the extra advantage of being able to leverage image features pre-trained from massive image datasets. As a matter of fact, we demonstrate in [5] that multi-view representations are still the state-of-the-art for shape classification tasks.

However, generating 3D shapes poses a more challenging situation. While generating 3D data, we are primarily concerned about generating surfaces, which are inherently sparse in the 3D space. This leads to a big drawback for occupancy grids: models using them require huge amounts of memory, being prohibitively large when generating high-resolution shapes. For multi-view representations, there are two main issues: first, these representations are restricted to representing only the visible portion of the surface – interior parts are not represented. Second, it is not clear how to enforce consistency between different views, which leads to a reduced quality in the generated shapes. A way to workaround the later is to enforce consistency through a post-processing optimization step, like the one we developed in [3]. Nevertheless, these models are still very memory intensive and do not handle generating multiple categories of objects.

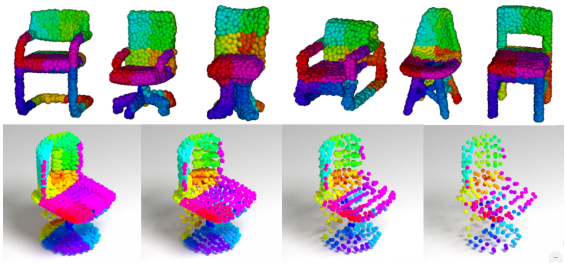

Figure 1: Point clouds sorted according to spatial partitioning structures induce reasonable point correspondence (indicated by similar colors). The same structure can also be used to compute multiple point cloud resolutions. We build upon these observations to design multi-resolution convolutional operators for point cloud data.

A reasonable alternative to those 3D representations is utilizing point clouds. Point clouds are a very compact surface representation – every point in the point must be part of the surface. They also naturally support extra surface attributes, like color and normals, and are directly captured by a variety of 3D sensors. The biggest challenge while using point clouds in deep networks lies in its unstructured nature. Since they are sets of points, point cloud representations need to be invariant to permutations. Moreover, differently form multi-view representations and occupancy grids, it is not clear what is the best way to use convolutional layers in point cloud data.

Our attempt in creating generative models for point clouds was bootstrapped by using spatial-partitioning data-structures to assign an approximate correspondence between points of different point clouds [2]. The motivation is simple: if one can induce such correspondence, point clouds can be treated as structured data. In practical terms, we compute a *kd*-tree for every point cloud and sort the points according to a level-order traversal in the leaves of the tree.

This sorting induces a reasonable correspondence between points, as shown in Figure 1. Using this correspondence, we compute a linear low-dimensional shape representation that were used to train the first Generative Adversarial Networks (GANs) for point cloud generation [2]. Later, we noticed that the spatial partitioning induces a local neighborhood that can be successfully used to define convolutional operations and to represent multiple point cloud resolutions [4]. We called these models Multi-Resolution Tree Networks (MRTNets) and applied them to a variety of discriminative and generative tasks, like shape classification, part segmentation, single-view reconstruction and VAEs. Some of the results are presented in Figure 2. The models have a small memory footprint when compared against multi-view and occupancy grids counterparts while yielding state-of-the-art results for point cloud classification, single-view reconstruction and unsupervised feature learning benchmarks.

Figure 2: Single-view reconstruction using MRTNets.

Learning 3D Shapes from Images

Images are much more prevalent than 3D data. The most used shape classification benchmark, ModelNet40, contains about 10 thousand shapes, whereas the most popular image classification benchmark, ImageNet, contains about 14 million images. Nevertheless, images usually contain real world entities which are inherently 3D. In other words, a lot of 3D information is encoded in images and being able to leverage that information to learn to generate 3D shapes is key to build models that can overcome the lack of 3D training data. We study this issue within a very challenging problem setup. Consider a set of silhouette images like the ones in Figure 3. Those images represent silhouettes of various objects from the same category. If we have viewpoint annotation and object identification (i.e. which images correspond to the same object) this problem can be easily solved using visual hull. We can make the problem harder by assuming that no viewpoint annotation is available. In that case, we can probably achieve a reasonable result by relying in Structure from Motion (SfM) techniques. The most difficult setup occurs when we neither have object identification nor viewpoint annotation. In this scenario, one can only rely on non-rigid SfM, which require a strong prior over the generated shapes. What happens when we have no information regarding 3D shapes? Can we still do something about it?

Our solution consists of utilizing deep generative models coupled with differentiable projection operators. The intuition is simple: given a dataset with images, we want to generate 3D shapes that, when projected into the image plane, will look like they came from the dataset. More precisely, we want to match the distribution of images in the dataset to the images created by projecting the generated 3D shapes. Fortunately, there is a class of deep learning models which is remarkably good in mimicing target image distributions: generative adversarial networks (GANs). Thus, we augment the GAN generator with a differentiable shape projection module which turns 3D shapes into silhouette images. The result is a 3D generative model that is trained without ever seeing any 3D data, only silhouettes of 3D objects. We name this model Projective GAN (PrGAN) [1]. Later, we extended this architecture to enable learning from extra image annotations (e.g. part segmentation) by designing projection modules capable of generating part-segmented images [7].

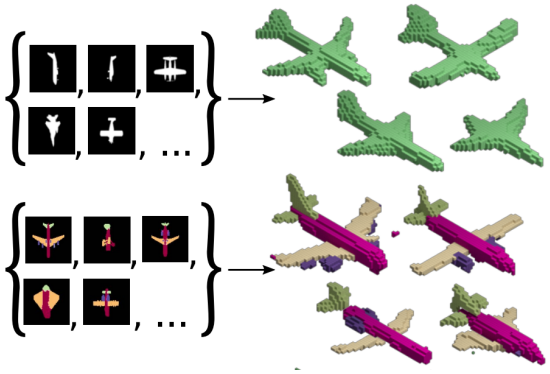

Figure 3: PrGAN is capable of learning generative models of 3D data without using any 3D supervision. The core of the approach is the utilization of differentiable projection operators that turn 3D representation into images of silhouettes and segmentation masks.

Another interesting scenario occurs when we don't have access to images of multiple objects, but only a small set images of a single object with viewpoint annotations. In that case, visual hull techniques are applicable, but we can do better than that, even without using any extra training data. We build upon priors induced by deep image models [6] and demonstrate that convolutional architectures can also induce priors over 3D shapes. We can then combine them with projection modules, deriving a class of powerful reconstruction techniques that does not rely on task-specific training. Additionally, by doing inference through descent we can update viewpoint estimations, enabling reconstruction with noisy viewpoint annotations. We also design new differentiable projection modules that enable learning 3D shapes from depth maps and sinograms, yielding state-of-the-art results for tomographic reconstruction [8] and visual hull.

Future Work

In my research, I investigated deep learning models for 3D data generation and understanding. We tackled two fundamental problems: models and data. Models for 3D data recognition and generation need to deal with the representation issue: differently from images, there are multiple 3D representations with different strengths and weaknesses in different tasks. We presented a technique to handle point cloud data that significantly outperformed well-established 3D representations in generative tasks while having a fraction of the memory requirements. On the other hand, 3D data itself is a scarce commodity. Our approach to this issue was to design differentiable operators that can turn 3D representations into images, inserting the image formation pipeline into deep learning architectures. This framework enabled using solely image supervision for generating 3D content in a variety of scenarios. Despite the progress made in those problems there is a lot left to be explored.

Learning 3D Shapes from Real Images. The projection modules developed in our research can be seen as simple differentiable renderers. The ultimate goal is not only be able to learn 3D from silhouettes but from photorealistic images. Computer graphics research has developed a solid toolset for synthesizing realistic images. Combining those techniques with good 3D shape representations will lead to better models, capable of learning not only 3D geometry but material properties, like color, textures and BRDFs. Models like these have many applications in image and 3D editing and are paramount to aid the creation of photorealistic content.

Generative Models for Shape Handles. Deep learning techniques have focused on generating 3D data using raw representations, like point clouds [4] and multi-view [3]. Despite achieving good results, those representations are not amenable to editing and using them to edit existing 3D shapes is not a trivial task. On the other hand, many geometry processing approaches have focused in designing algorithms to summarize complex shapes in simpler structures that can be used for shape editing, which can be referred to as shape handles. Deep learning models capable of generating those representations is a fundamental step in doing for 3D data what deep generative models have done for many image domains: enable high quality synthesis with minimal user proficiency.

Deep Priors for 3D Shapes. As shown in [8], convolutions induce good priors for occupancy grids, but their limitation in learning-based scenarios indicate that there are more efficient approaches, like recent methods based on surface parametrizations or implicit functions. While some work has been done using those models with task-specific training, the behavior of them as priors remains unexplored. This line of work is particularly useful for aiding creative 3D applications, since it does not require any training data and therefore is not restricted to models trained on limited 3D datasets.

Single-view Scene Reconstruction. Techniques that yield the best single-view reconstruction results are usually applied in images containing a single object. While a lot of progress has been made in this setup, it is time to move on to more interesting and realistic scenarios. A natural extension is to build upon consolidated object detection pipelines and utilize single-object reconstruction models in the process of reconstructing full scenes containing multiple objects. However, such model should not only be able to reconstruct 3D from single-object proposals but also utilize contextual scene information to guide its reconstruction.

References

[1] Matheus Gadelha, Subhransu Maji, and Rui Wang. "3D Shape Induction from 2D Views of Multiple Objects". In: *International Conference on 3D Vision (3DV)* (2017).

[2] Matheus Gadelha, Subhransu Maji, and Rui Wang. "Shape Generation using Spatially Ordered Point Clouds". In: *British Machine Vision Conference (BMVC)*. Sept. 2017.

[3] Zhaoliang Lun, Matheus Gadelha, Evangelos Kalogerakis, Subhransu Maji, and Rui Wang. "3D Shape Reconstruction from Sketches via Multi-view Convolutional Networks". In: *2017 International Conference on 3D Vision (3DV)*. 2017.

[4] Matheus Gadelha, Rui Wang, and Subhransu Maji. "Multiresolution Tree Networks for 3D Point Cloud Processing". In: *The European Conference on Computer Vision (ECCV)*. Sept. 2018.

[5] Jong-Chyi Su, Matheus Gadelha, Rui Wang, and Subhransu Maji. "A Deeper Look at 3D Shape Classifiers". In: *Second Workshop on 3D Reconstruction Meets Semantics (ECCV)*. 2018.

[6] Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. "A Bayesian Perspective on the Deep Image Prior". In: *Computer Vision and Pattern Recognition (CVPR)*. 2019.

[7] Matheus Gadelha, Aartika Rai, Subhransu Maji, and Rui Wang. "Inferring 3D Shapes from Image Collections using Adversarial Networks". In: *(In submission to IJCV)* (2019). arXiv: 1906.04910. URL: http://arxiv.org/abs/1906.04910.

[8] Matheus Gadelha, Rui Wang, and Subhransu Maji. "Shape Reconstruction using Differentiable Projections and Deep Priors". In: *International Conference on Computer Vision (ICCV)*. 2019.