In recent years, Machine Learning techniques have revolutionized solutions to longstanding 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 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], recognizing [5, 4, 10] and editing [8] 3D data? Is it possible to leverage image data to build models capable of reasoning about the world in 3D [7, 1, 9]?

Our research findings show that we are able to build models that efficiently generate 3D shapes as unstructured data – sets of geometric entities like points or shape handles. Those models require significantly less memory while generating higher quality shapes than the ones based on voxels and multi-view representations. For point clouds, our models are applied to multiple tasks: shape classification, segmentation, single-view reconstruction and unsupervised representation learning with variational auto-encoders (VAEs). However, those point cloud models are not very helpful if the user wants to use them for shape editing; differently from images, where rasterized representations provide intuitive editing operations, users tend to have a lot of difficulty editing raw 3D representations like point clouds and occupancy volumes. We address this issue by proposing a class of deep generative models for shape handles, *i.e.* geometric proxies that approximate an underlying 3D shape while being amenable to user editing. Notwithstanding, all aforementioned techniques require explicit 3D supervision, which is scarce. My research also investigates ways of learning with limited 3D data for both discriminative and generative models. For point cloud recognition tasks, we proposed a self-supervised task based on classic geometric decomposition methods that yield state-of-the-art results for few-shot part segmentation and unsupervised classification tasks. For generative models, we propose to deal with the lack of 3D data by leveraging images and investigating the shape priors induced by various deep networks. In the core of these approaches are 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 with Irregularly Structured 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 on massive image datasets. As a matter of fact, we demonstrate in [5] that multi-view representations are still a very competitive baseline, surpassing more recent 3D classification approaches.

However, the situation is quite different when we are trying to generate 3D shapes. 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.

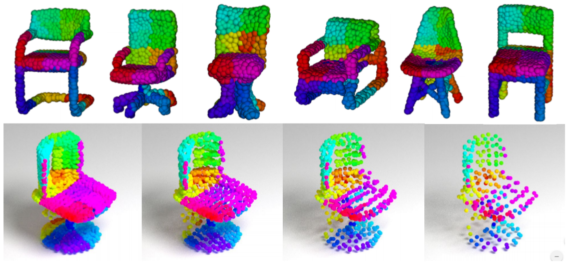

Figure 1: Point clouds sorted according to spatial partitioning structures induce reasonable point correspondence (indicated by similar colors). The same structure can also be used to compute multiple point cloud resolutions. We build upon these observations to design multi-resolution convolutional operators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

Point Clouds. A reasonable alternative to those 3D representations is utilizing point clouds. Point clouds are a compact surface representation – every 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 irregular nature – they are not defined in a regular grid. 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.

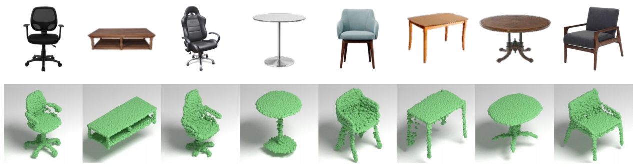

Figure 2: Single-view reconstruction using MRTNets. More info here.

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 rough correspondence between points, as shown in Figure 1. Using this correspondence, we compute a linear low-dimensional shape representation that we 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 with multi-view and occupancy grids counterparts while yielding state-of-the-art results in point cloud classification, single-view reconstruction and unsupervised feature learning benchmarks.

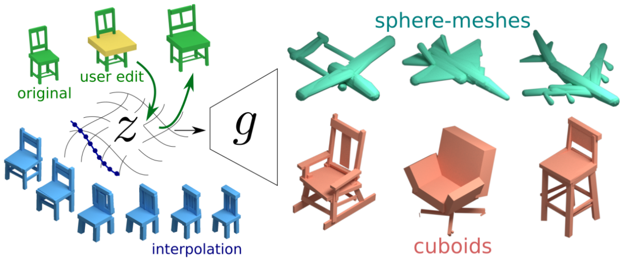

Figure 3: Generative models for shape handles. Our class of models is capable of dealing with multiple types of shape handles and is applicable to multiple tasks. More info here.

Shape handles. While using point clouds for shape recognition and generation proved to be reasonably effective, editing shapes represented as point clouds is a very laborious task. Unfortunately, the same is also true for other representations used in deep generative models, like multi-view, occupancy grids and implicit functions. However, there are years of research within the geometric processing community concerning *shape handles* – geometric proxies that approximate a more complex underlying shape while being amenable to user manipulation. Our solution is a connection between classical geometric processing and modern machine learning – we propose a class of deep generative models capable of generating shapes represented as a set of shape handles [8]. The key idea is to apply a set-to-set reconstruction losses on sets of shape handles while deriving a simple and versatile mechanism to compare pairs of shape handles. We show that, as long one is capable of computing a point-to-handle distance, our method is capable of generating different types of shape handles, like cuboids, ellipsoids and even sphere-meshes. The models can then be used for shape parsing, interpolation or editing, while handling generating sets of variable cardinality in a single forward pass.

Learning with Limited 3D Supervision

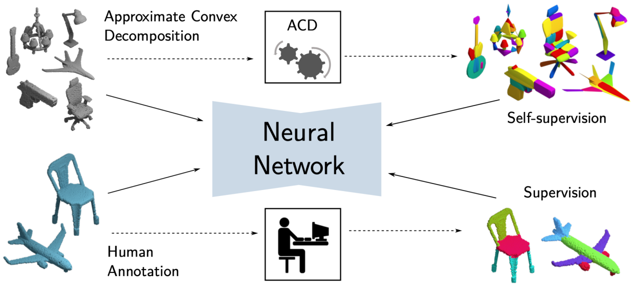

Figure 4: Point cloud representation learning using ACD. Our self-supervised learning task leverages large unlabelled shape repositories to learn better point embeddings for classification and segmentation. More info here.

An important part of my research concerns developing methods to reason about the 3D space while very limited 3D data is available (or none altogether). We start by investigating how to train point cloud recognition models when point labels are scarce. For example, if we want to train a model to segment the 3D scan of an airplane into parts like wings, engine and fuselage, we need to rely on massive amounts of *annotated* point clouds – every point of every airplane needs to be labeled as belonging to one of this parts. This requires a huge amount of human effort, which considerably increases the cost of creating such models. On the other hand, there are big unlabeled shape repositories that are usually not leveraged for those tasks. How can we use all this unlabeled 3D data to learn better models for 3D recognition. We tackle this problem by designing a self-supervised task based on Approximate Convex Decomposition (ACD). ACD is a classical shape decomposition technique that breaks down a shape into a set of approximately convex parts. Our main observation is the fact that many part segmentation boundaries correspond to the ones computed by ACD. In other words, if we can teach a network to perform ACD, its features are probably useful for downstream tasks like shape classification and part-segmentation. And, since ACD can be automatically computed for any 3D shape, we can generate part labels for all the shapes in large shape repositories and use them to train point cloud recognition models [10]. We demonstrate that models trained in this manner achieve state-of-the-art performance in few-shot part segmentation and unsupervised shape classification benchmarks.

Learning from images. Even though scanning technologies and shape repositories are increasingly more popular, the simple fact is that 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, pictures are projections of 3D content in the image plane. 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 propose to study this issue within a very challenging problem setup. Consider a set of silhouette images like the ones in Figure 5. 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 [9].

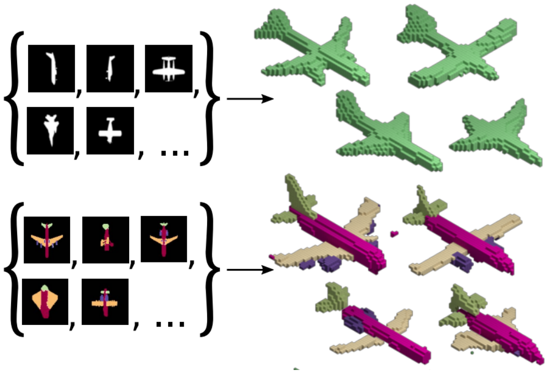

Figure 5: PrGAN is capable of learning generative models of 3D data without using any 3D supervision. The core of the approach is the utilization of differentiable projection operators that turn 3D representation into images of silhouettes and segmentation masks. More info here.

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 gradient 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 training-free tomographic reconstruction [7] and visual hull.

Future Work

In my research, I investigated deep learning models for 3D data generation and understanding. I focused on 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. My work explored multiple combinations of those. For shape recognition and reconstruction, we presented architectures for dealing with point clouds. For shape editing and manipulation, we proposed a class of models capable of directly generating shape handles. Both of these representations fall under the broader category of unstructured (or set-based) representations and, in comparison to the more common multi-view and occupancy grid counterparts, show remarkable memory and computational efficiency. The other big issue concerning learning models for 3D reasoning is data itself. 3D data is a scarce commodity. When 3D data exists, but it is unlabeled, we proposed a self-supervised task based on approximate convex decomposition. When 3D data itself is not available, but we have access to images, we proposed techniques relying on differentiable projection operators. 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 from real images requires more than geometry. 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, within a single framework. Models like these have many applications in image and 3D editing and are paramount to aid the creation of photorealistic content and 3D manipulation of 2D images. Perhaps more importantly, techniques capable of reasoning about various components of the image formation process can lead to better understanding of the inductive biases needed to perform more efficient learning in various scenarios. In other words, how can one design models that induce good priors for other scene components besides geometry, like materials and illumination?

Models for Dynamic 3D Data. The main motivation of my research comes from the intuition that reasoning with images is suboptimal when interacting with the real tridimensional world. However, for many applications, considering a static 3D world is also a crude approximation. After significant progress in developing models for static objects, a natural next step is to think about problems concerning dynamic 3D scenes. This requires rethinking many decisions concerning architectures and representations for 3D data – a solution that simply adds an extra dimension is clearly not the best one. More than that, foundational work creating datasets for these types of problems is required. Models that are capable of reasoning about motion in 3D already have a clear impact in autonomous navigation. Models that can generate moving data in 3D will likely not only have a similar impact in creative applications, but may also lead to efficient 3D video representations.

Full-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. Whereas seminal work has been done in this area, there is still a significant gap between the quality of reconstructing single objects when compared to full scenes. Closing this gap requires investigating efficient 3D representations and building datasets with full 3D annotation. Reconstructing high-quality full 3D scenes from as few images as possible (maybe just one) is a key component for developing the next generation of robots and augmented reality applications.

Differentiable Programming for Computer Aided Design. Using differentiable rendering to obtain supervisory signal for learning 3D shapes has shown remarkable success. The main intuition behind those approaches is that, since we know how to simulate the image formation process, one can devise differentiable approximations and it to the model training loop. However, there are many other domains where such approaches can be applied. For example, if one can differentiably simulate agent behavior and interaction with objects, the signal can be used to guide many design tasks. More than that: artists and game designers often rely on graph-based descriptions of materials, shaders and game logic. Implementing differentiable approximations to these workflows would allow learning from data more efficiently and recovering interpretable latent representations, crucial for allowing meaningful manipulation by practitioners. These differentiable engines have the potential to become the cornerstone of accelerated content creation workflows for games, visual effects and 3D design in general.

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: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

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

[8] Matheus Gadelha, Giorgio Gori, Duygu Ceylan, Radomir Mech, Nathan Carr, Tamy Boubekeur, Rui Wang, and Subhransu Maji. "Learning Generative Models of Shape Handles". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[9] Matheus Gadelha, Aartika Rai, Subhransu Maji, and Rui Wang. "Inferring 3D Shapes from Image Collections using Adversarial Networks". In: *International Journal of Computer Vision (IJCV)* (2020).

[10] Matheus Gadelha*, Aruni RoyChowdhury*, Gopal Sharma, Evangelos Kalogerakis, Liangliang Cao, Erik Learned-Miller, Rui Wang, and Subhransu Maji. "Label-Efficient Learning on Point Clouds using Approximate Convex Decompositions". In: *European Conference on Computer Vision (ECCV)*. 2020.