

Towards More Physically Plausible Generative Models

CVPR 2025 Workshop:

Ind3D: *Enforcing Inductive Bias in 3D Generation from Geometric, Physical, Topological, and Functional Perspectives*

Maks Ovsjanikov

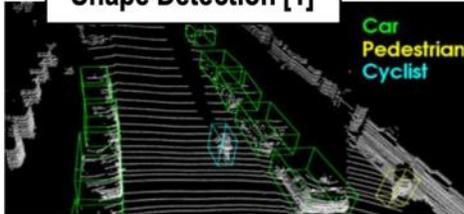
Joint work with: *M. Mezghanni,, M. Boulkenafed, L. Maillard, T. Durand, ...*



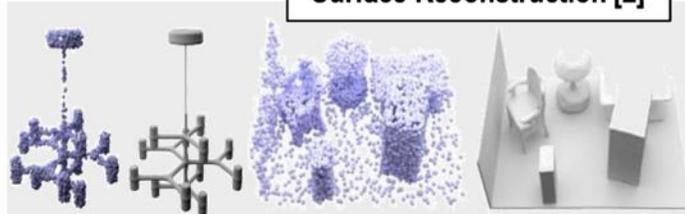
Motivation: Gap Between Methodology and Applications

Methods

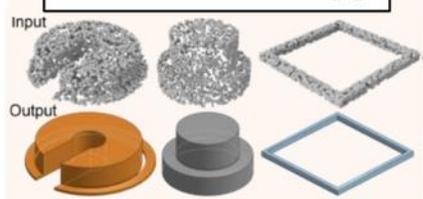
Shape Detection [1]



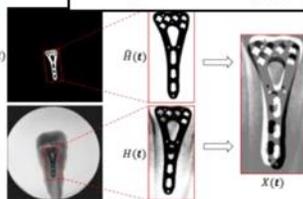
Surface Reconstruction [2]



CAD Reconstruction [3]



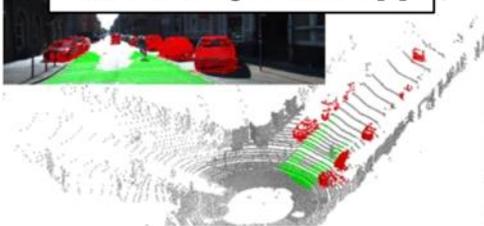
Registration [4]



3D Clothing [5]



Semantic Segmentation [6]



Generative Modeling [7]



Real-world applications

3D Modeling



Entertainment



Medicine



Robotics



Autonomous Driving



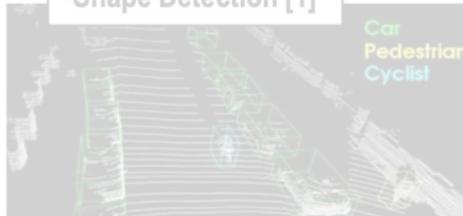
AR/VR



Motivation: Gap Between Methodology and Applications

Methods

Shape Detection [1]



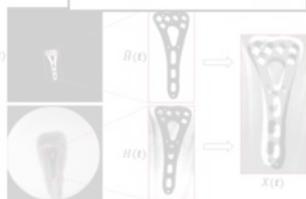
Surface Reconstruction [2]



CAD Reconstruction [3]



Registration [4]



3D Clothing [5]



Semantic Segmentation [6]



Generative Modeling [7]



Real-world applications

3D Modeling



Entertainment



Medicine



Robotics



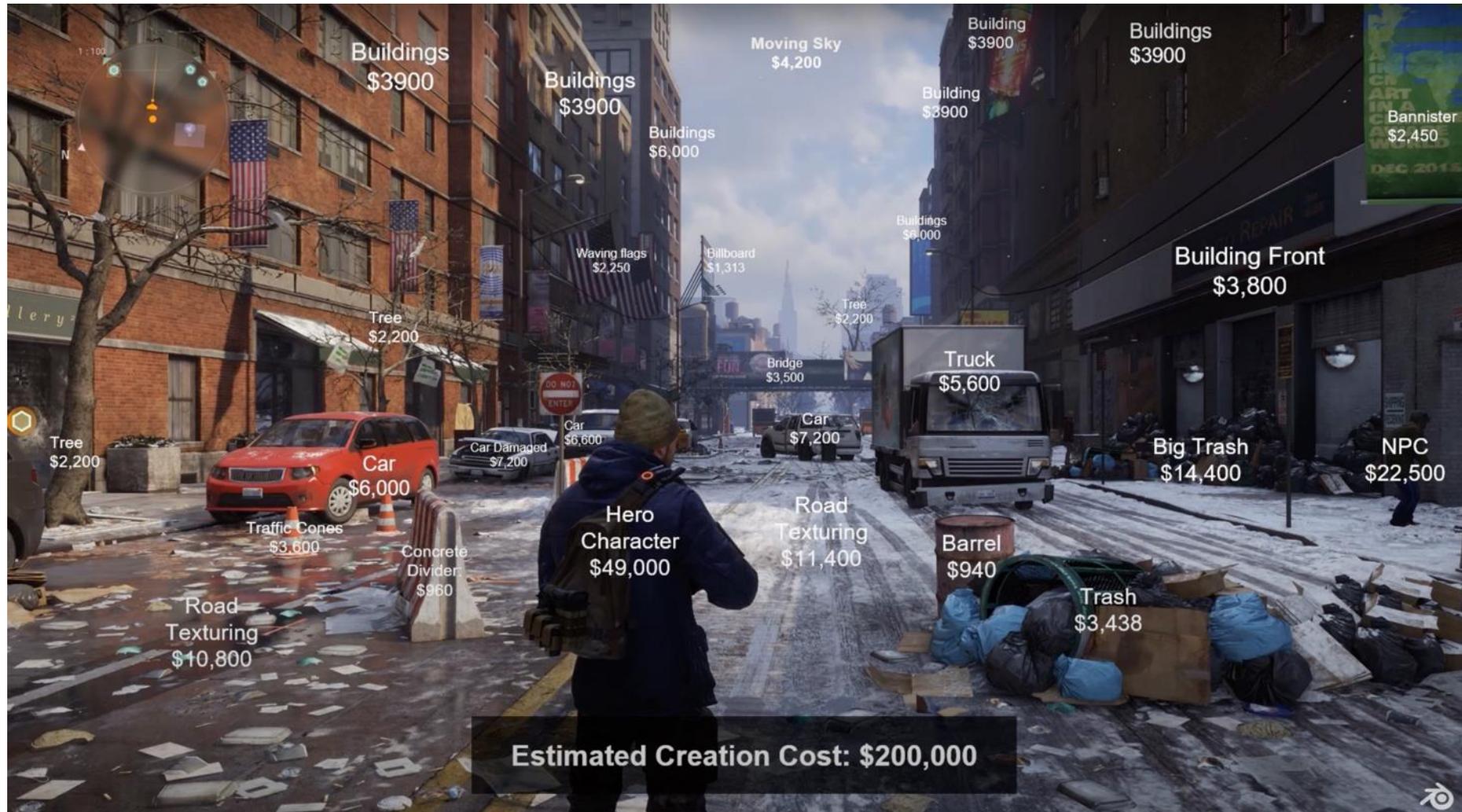
Autonomous Driving



AR/VR



3D Modeling is Expensive



3D Generative Modeling Goals

Realism

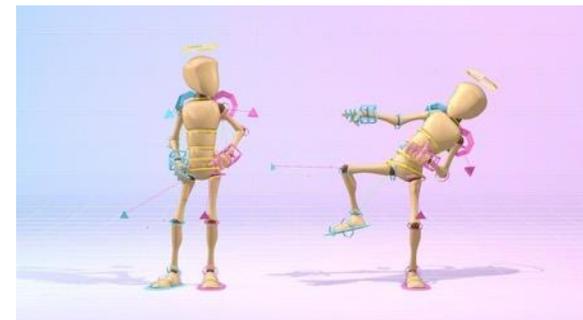
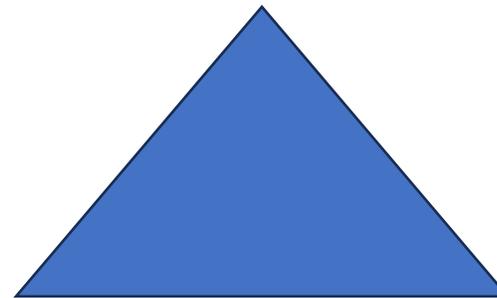


Image source:
Evermotion



Image source: Objaverse-XL

**Data
Efficiency**



Source: Ugur
Ulvi Yetiskin

Controllability

Plan for Today

Two major topics:

1. Generative Modeling for 3D Objects
2. Generative Modeling for indoor 3D Scenes

Mariem Mezghanni



Léopold Maillard



Physically-aware Generative Network for 3D Shape Modeling

Mariem Mezghanni¹ Maïka Bouhadef¹ André Lantier¹ Maks Ovsjanikov^{1*}

¹LIX, Ecole Polytechnique, IP Paris²
mezghanni.maïka@lix.polytechnique.fr

Abstract

Shapes are often designed to satisfy structural properties and serve a particular functionality in the physical world. Unfortunately, most existing generative models focus primarily on the geometry or visual plausibility, ignoring the physical or structural constraints. To remedy this, we propose a novel method aimed to induce deep generative models with physical awareness. In particular, we introduce a loss and a learning framework that promote the characteristics of the generated shapes: their connectivity and physical stability. The former ensures that each generated shape consists of a single connected component, while the latter promotes the stability of the shape when subjected to gravity. Our proposed physical losses are fully differentiable and we demonstrate their use in end-to-end learning. Crucially, we demonstrate that such physical objectives can be achieved without sacrificing the expressive power of the model and stability of the generated models, the demonstration through extensive comparisons with the state-of-the-art deep generative models, the ability and efficiency of our proposed approach, which avoiding the previously noted differentiable physical constraints at training time.

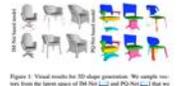


Figure 1: Visual results for 3D shape generation. We sample vectors from the latent space of the first and second PCA axes that we denote using the corresponding feature vectors (first row) and use generative network trained with the proposed physical losses (second row). Performance metrics are marked to red circle. The resulting shapes become more connected and physically viable.

Shapes are also expected to satisfy physical and functional constraints. Consequently, the generated content might appear to be a convincing example of a particular category (e.g., a chair, a car etc.) but there is no guarantee that it can be feasible and functional in the physical world. There has been a steady stream of works in the design community in studying 3D shapes from a functional perspective. Last but, previous attempts in developing generative neural networks for representing functional and structural 3D shapes have not yet jointly leveraged the power of analyzing generative, physical and functional representations. Although it seems relatively straightforward for a human designer to make explicit connections between geometry, physics and functionality, it is still challenging to train intelligent models to do the same.

In this paper, we introduce a physically-aware generative modeling method that makes a step to overcome these limitations (Fig. 1). We work a latent representation that incorporates generative, structural and physical information. Such a latent space enables many new physical applications including generating novel and realistic shapes, physical shape optimization, etc. To this end, we introduce a loss that induces training deep generative models of 3D shapes with physical awareness.

CVPR 2021

Physical Simulation Layer for Accurate 3D Modeling

Mariem Mezghanni¹ Théo Bodrozic¹ Maïka Bouhadef¹ Maks Ovsjanikov^{1*}

¹LIX, Ecole Polytechnique, IP Paris
mezghanni.maïka@lix.polytechnique.fr

Abstract

We introduce a novel approach for generative 3D modeling that explicitly encompasses the physical and structural constraints of the generated shapes. To this end, we advocate the use of online physical simulation as part of learning a generative model. Unlike previous neural methods, our approach is trained end-to-end and with a fully differentiable physical simulation in the training loop. We accomplish this by leveraging recent advances in differentiable programming, and introducing a fully differentiable general physical simulation layer, which accurately evaluates the shape's stability when subjected to gravity. We then incorporate this layer as a regular distance function (DDF) shape details. By incorporating a continuous DDF decoder with our simulation layer, we demonstrate through extensive experiments that online physical simulation improves the accuracy, visual plausibility and physical stability of the resulting shapes, while requiring no additional data at an training stage.



Figure 1: Qualitative comparison of online simulation with the back-offer simulation (LL) for the task of shape optimization. From top to bottom: physically viable shapes sampled from back-offer DDF; results using LL and our method. The optimized shapes reflect the accuracy and efficiency of online simulation compared to the offline setting in terms of physical quality and generative consistency.

1. Introduction

Over the past several years, there has been a steady stream of works aimed at developing deep neural networks for 3D shape generation. A key challenge is to accurately describe plausible and diverse content while preserving geometry and structural stability (Liu et al., 2020). Though remarkable progress has been made in this direction, most of the state-of-the-art approaches primarily on generative or visual plausibility, while overlooking a key property of 3D design: functionality (Fig. 1). Indeed, generating 3D shapes often means to serve a particular function in the real world. For instance, a chair is expected to be stable when subjected to gravity. Ignoring this crucial constraint hampers the generated content to reflect the same functional artifacts such as lack of connectivity or physical instability (LL), which severely hinders its utility in real-world downstream tasks.

One way to address this challenge is by leveraging phys-

CVPR 2022 (Oral)

DeBaRA: Denoising-Based 3D Room Arrangement Generation

Léopold Maillard^{1,2} Nicolas Serrurier-Garcos¹ Tom Durand¹ Maks Ovsjanikov^{1*}

¹LIX, Ecole Polytechnique, IP Paris ²Thussati Systèmes
maillard.leopold@lix.polytechnique.fr {tsurand, harsman}@thussati.com

Abstract

Generating realistic and diverse layouts of furnished indoor 3D scenes unlocks multiple interactive applications impacting a wide range of industries. The inherent complexity of object interactions, the limited amount of available data and the requirement to fulfill spatial constraints all make generative modeling for 3D scene synthesis and arrangement challenging. Current methods address these challenges autoregressively or by using off-the-shelf diffusion objectives by simultaneously predicting all attributes without 3D reasoning considerations. In this paper, we introduce DeBaRA, a score-based model specifically tailored for precise, controllable and flexible arrangement generation in a bounded environment. We argue that the most critical component of a scene synthesis system is to accurately establish the size and position of various objects within a restricted area. Based on this insight, we propose a lightweight conditional score-based model designed with 3D spatial awareness at its core. We demonstrate that by focusing on spatial attributes of objects, a single trained DeBaRA model can be leveraged at test time to perform several downstream applications such as scene synthesis, completion and re-arrangement. Further, we introduce a novel Self Score Evaluation procedure so it can be optimally employed alongside external LLM models. We evaluate our approach through extensive experiments and demonstrate significant improvement upon state-of-the-art approaches in a range of scenarios.

1 Introduction

Systems capable of generating realistic environments comprising multiple interacting objects would impact several industries including video games, robotics, augmented and virtual reality (AR/VR) and computer-aided interior design. As a result and in tandem with the growing availability of synthetic datasets of indoor layouts (Li et al., 2023), which can be populated with high-quality 3D assets (Li et al., 2023), data-driven approaches for automatically generating and arranging 3D scenes have been actively investigated by the computer vision community. Notably, the ongoing success of deep generative models for controllable content creation in the text and image domains has recently been extended to scene synthesis, allowing users to craft realistic indoor environments from a set of multimodal constraints (Yu, 2023; Yu et al., 2023).

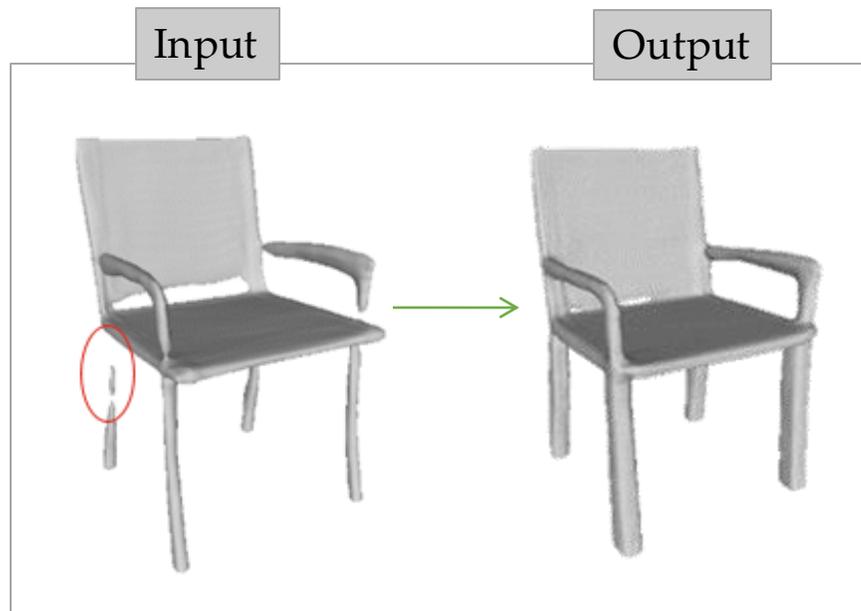
Challenges associated with 3D indoor scene generation are numerous as the intricate nature of multi-object interactions is difficult to capture and model precisely. Items should be placed, prioritarily oriented and oriented relative to one another, in a way that is both plausible and aligned with subjective and context-dependent priors such as style, as well as ergonomic and functional preferences. Additionally, objects should fit within a bounded, restricted area, and a viable arrangement can break the perceived validity of the synthesized environment (e.g., overlapping, floating or out-of-bounds

*Work done during internship at Thussati Systèmes.

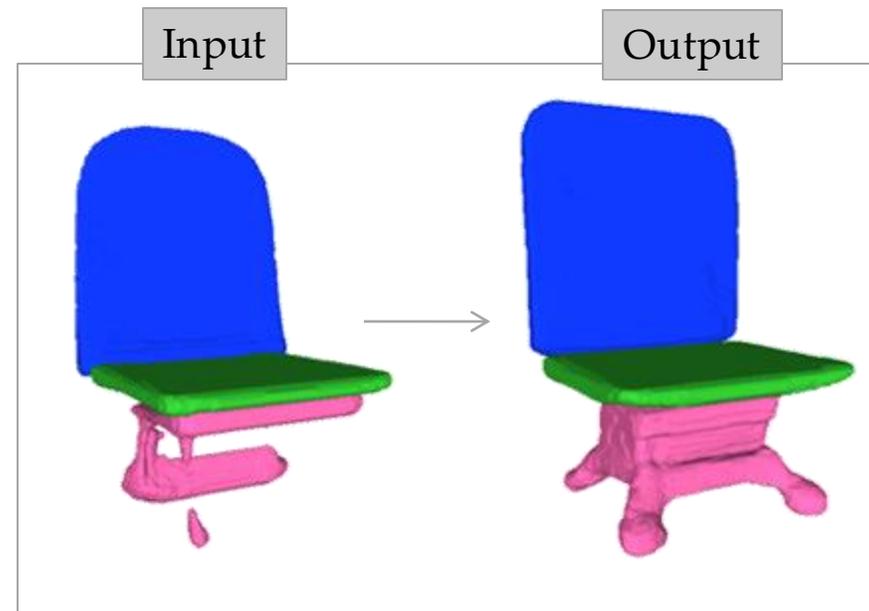
NeurIPS 2024

Overall Goal

Endow generative networks biases to promote **connectivity** and **physical stability**.



Connectivity



Physical stability

M. Mezghanni, et al. "Physically-aware generative network for 3d shape modelling," CVPR 2021.

M. Mezghanni, et al. "Physical simulation layer for accurate 3d modeling," CVPR 2022.

Why connectivity and physical stability ?

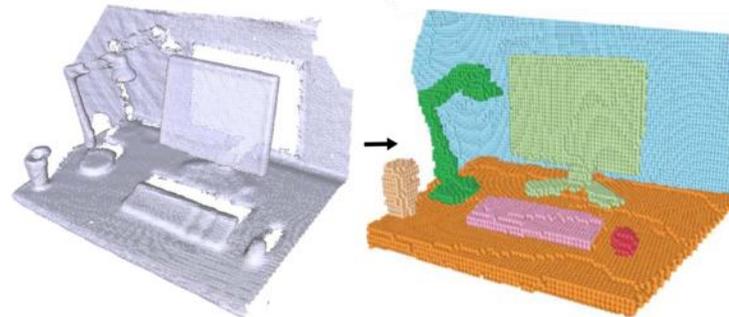
- Frequent cause of failures
- Represent a shared functional requirement across different shape categories
- Physical stability has proved beneficial for boosting many computer vision and graphics tasks.



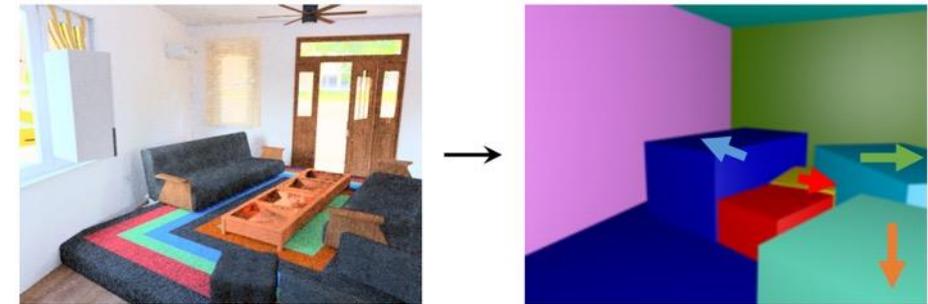
Example functional failures



3D printing [1]



Scene segmentation [2]



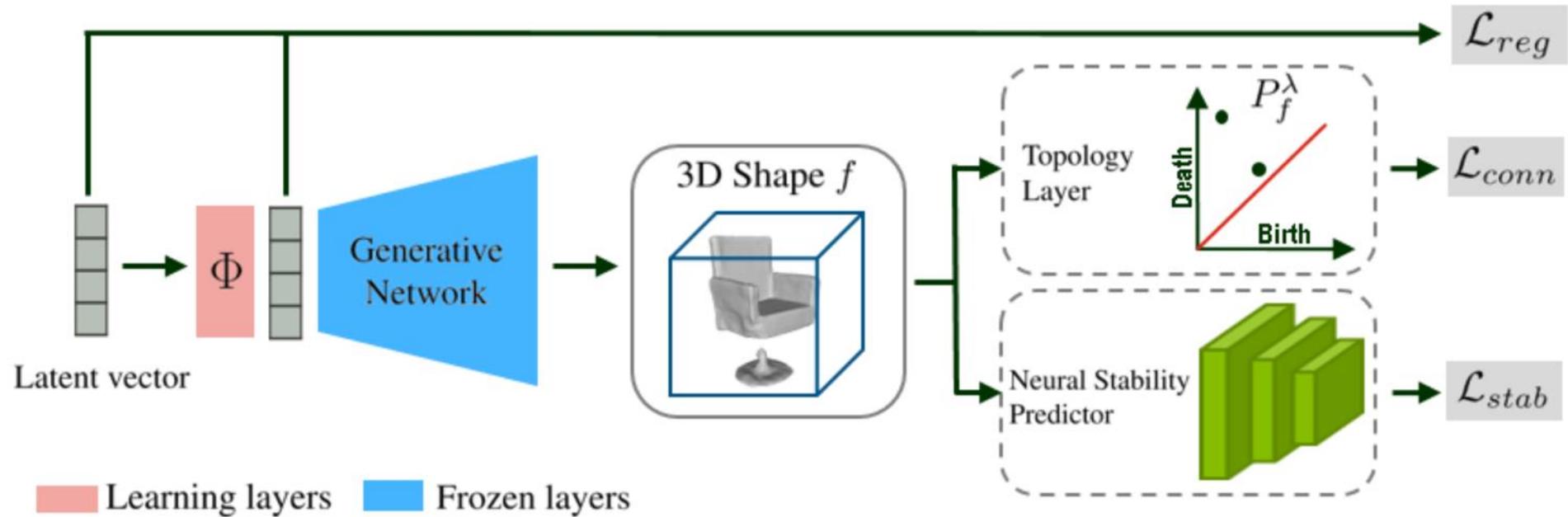
3D reconstruction [3]

[1] Make It Stand: Balancing shapes for 3D fabrication, Prévost et al., ACM SIGGRAPH, 2013.

[2] Beyond point clouds: Scene understanding by reasoning geometry and physics? Zheng et al., CVPR, 2013

[3] Learning to exploit stability for 3d scene parsing.. Du et al., NeurIPS, 2018

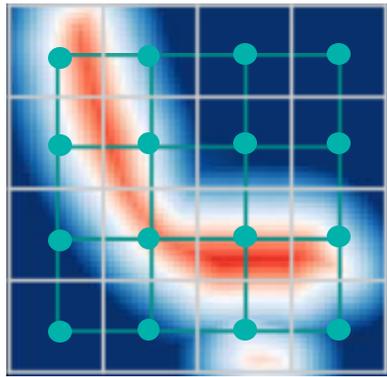
Method – Overview



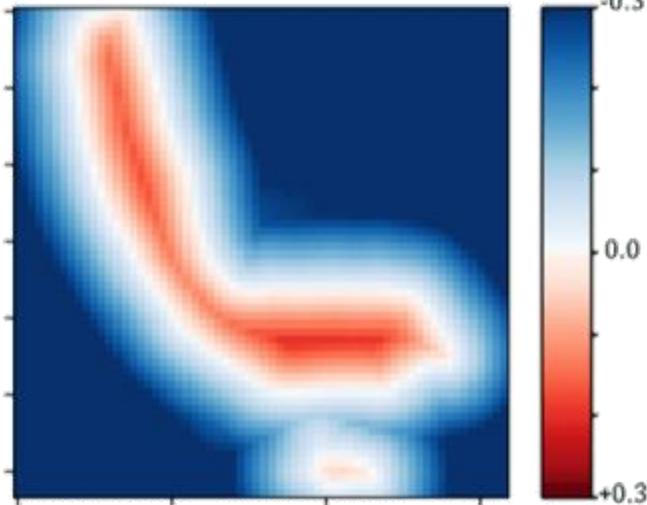
Method

Differentiable Connectivity Loss via Persistent Homology

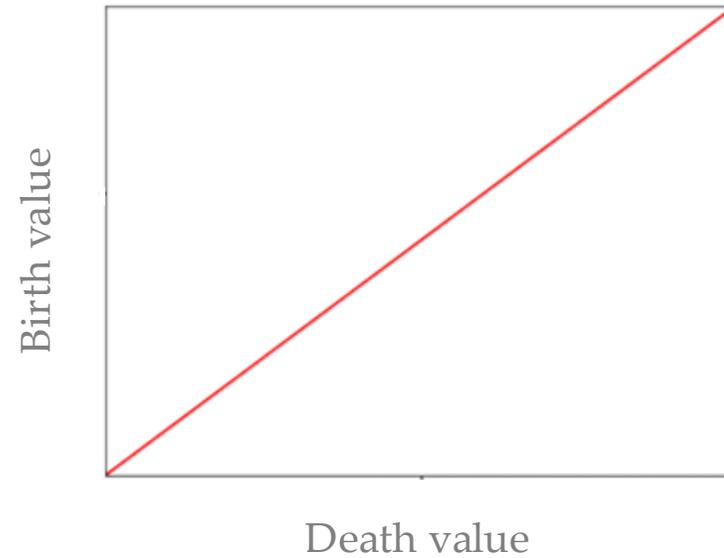
f : an implicit function



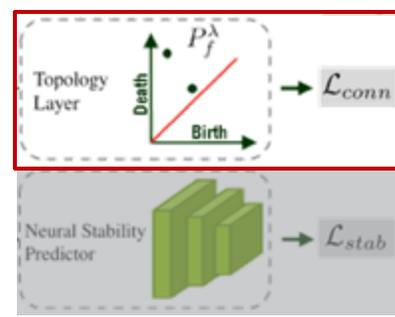
Cubical Complex



Persistence diagram P_f^λ



λ -isosurface extraction

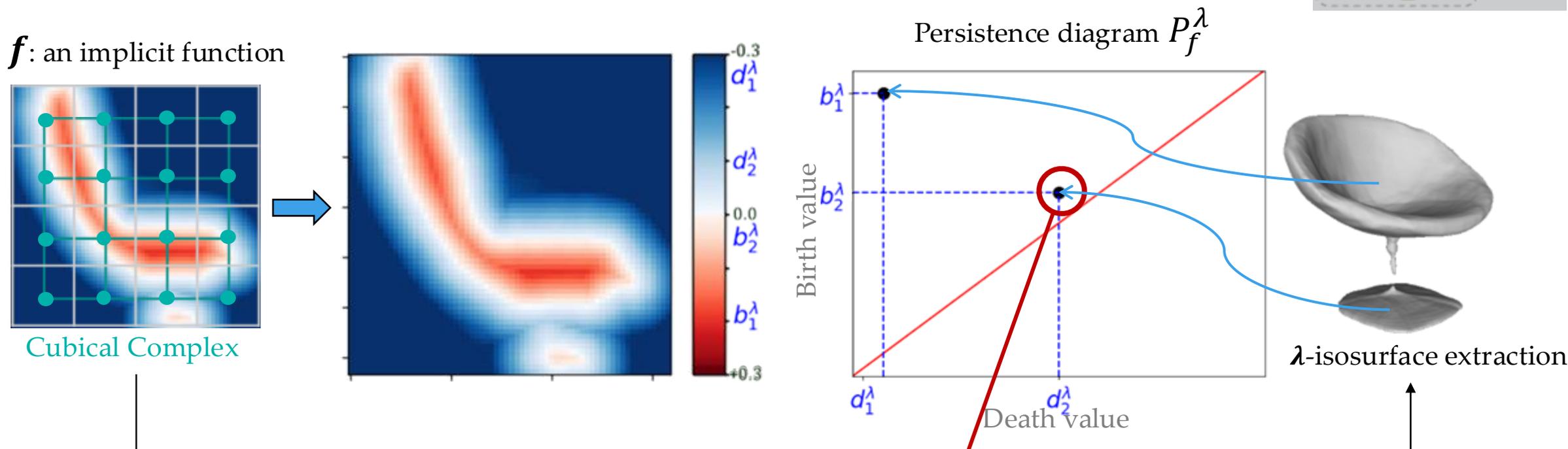


[1] A topology layer for machine learning, Gabrielsson et al., PMLR, 2020

[2] Topological Function Optimization for Continuous Shape Matching, Poulenard et al., CGF, 2018

Method

Differentiable Connectivity Loss via Persistent Homology



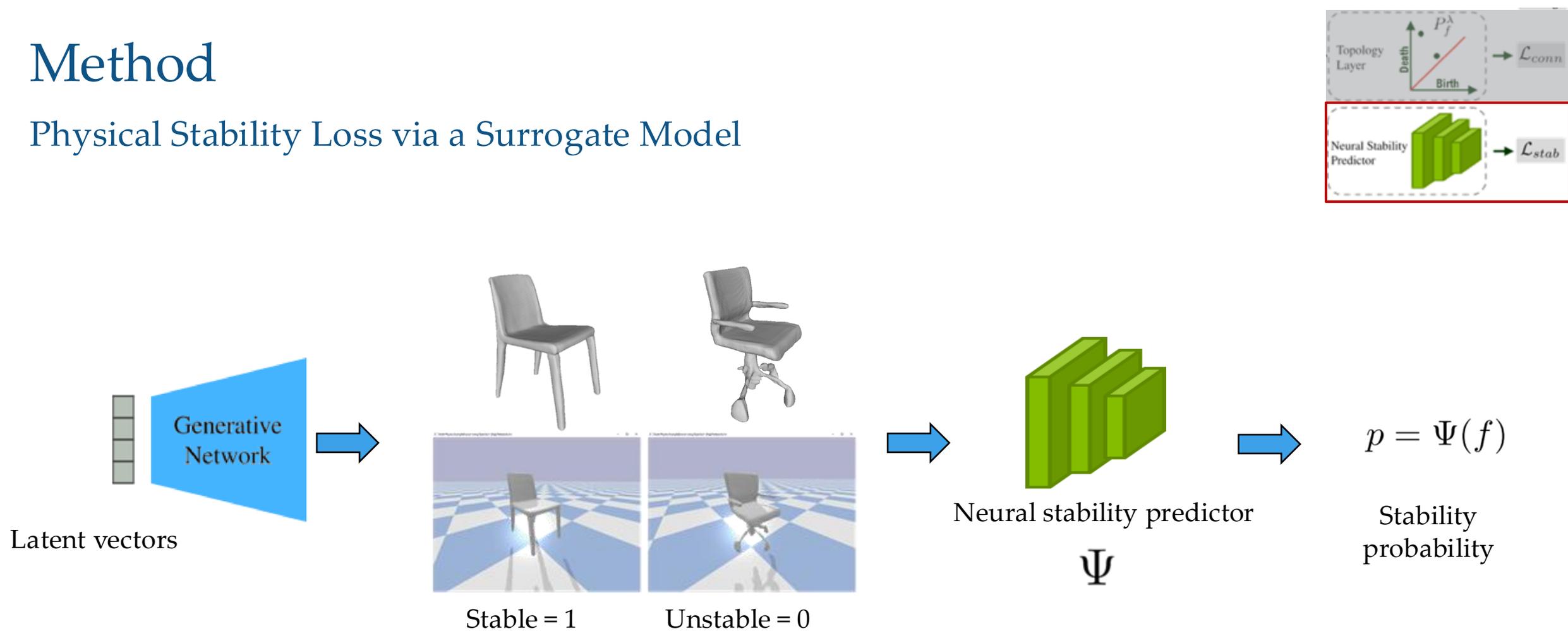
$$\mathcal{L}_{conn} = \sum_{(b_i^\lambda, d_i^\lambda) \in P_f^\lambda; \geq 2} (b_i^\lambda - d_i^\lambda)$$

[1] A topology layer for machine learning, Gabrielsson et al., PMLR, 2020

[2] Topological Function Optimization for Continuous Shape Matching, Poulenard et al., CGF, 2018

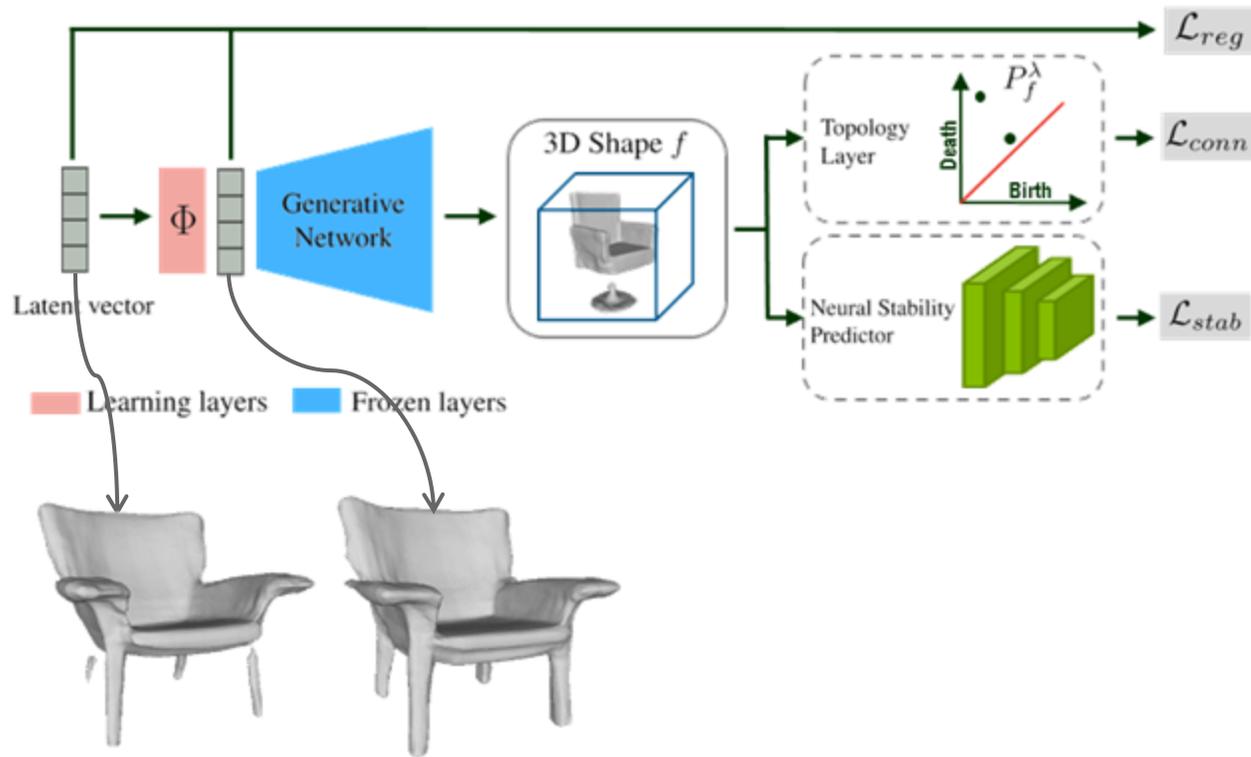
Method

Physical Stability Loss via a Surrogate Model



$$\mathcal{L}_{stab} = \max(1 - \Psi(f), \alpha); \alpha = 0.5$$

Learning Framework



Training stages:

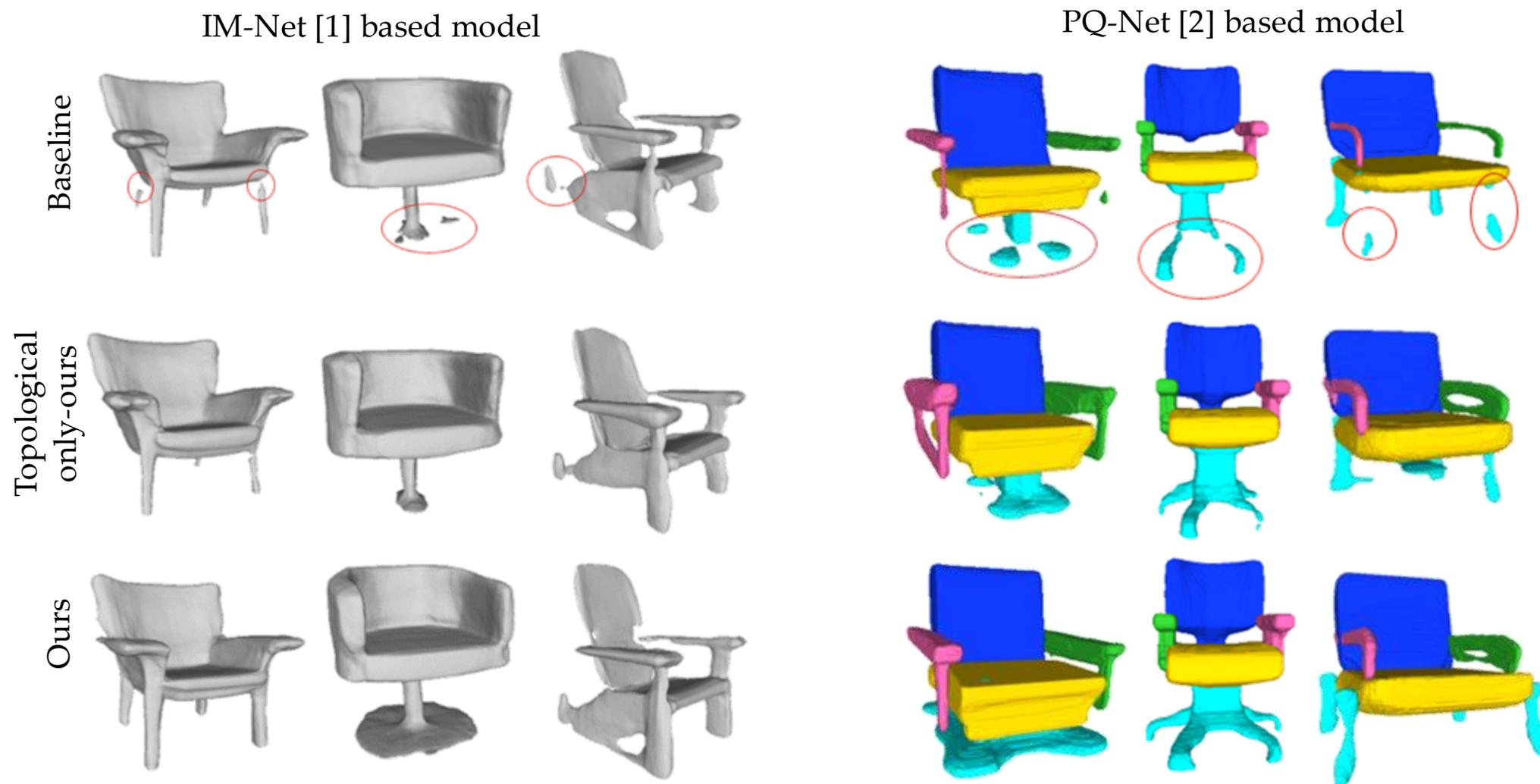
1. Train a generative network G
2. Freeze G and train a **mapping network** Φ

Motivation: preserve the diversity and quality of the generated content since the latent space of objects is unchanged.

$$\mathcal{L}_{reg}(z) = \|z - \Phi(z)\|_2; \quad \mathcal{L}_{total} = \mathbb{E}_{z \in \mathcal{V}} [\mathcal{L}_{reg} + \alpha_c \mathcal{L}_{conn} + \alpha_s \mathcal{L}_{stab}]$$

Learned latent space of shapes

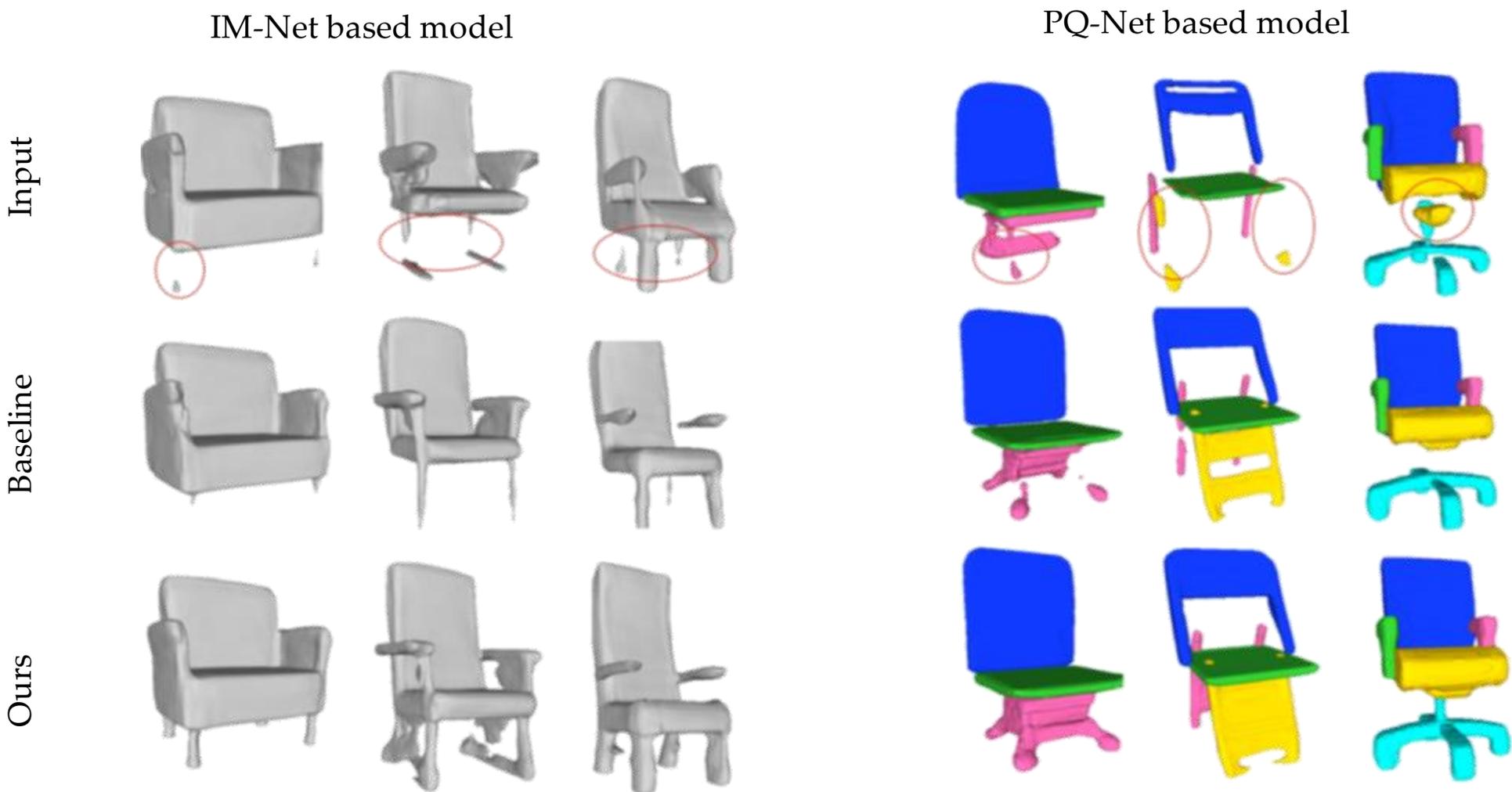
Results: Shape Generation



[1] IM-NET: Learning implicit fields for generative shape modeling. Chen et al., CVPR, 2019

[2] PQ-NET: A generative part Seq2Seq network for 3D shapes. Wu et al., CVPR, 2020

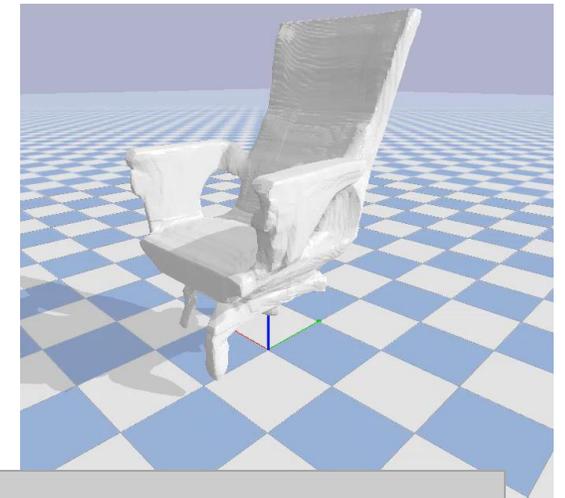
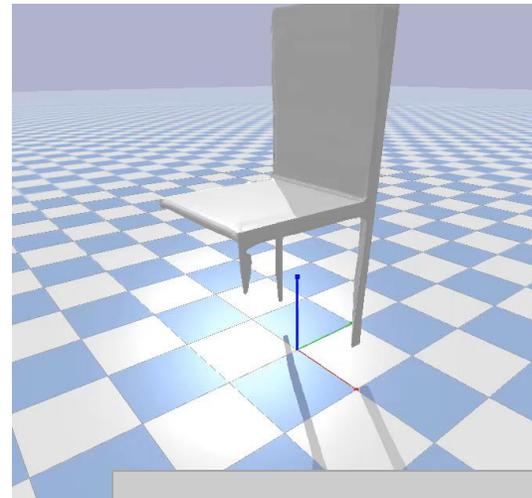
Results: Shape Correction



Offline vs Online Simulation

Offline simulators

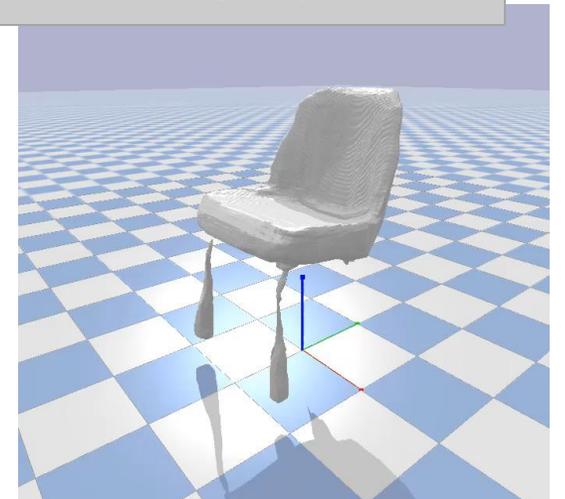
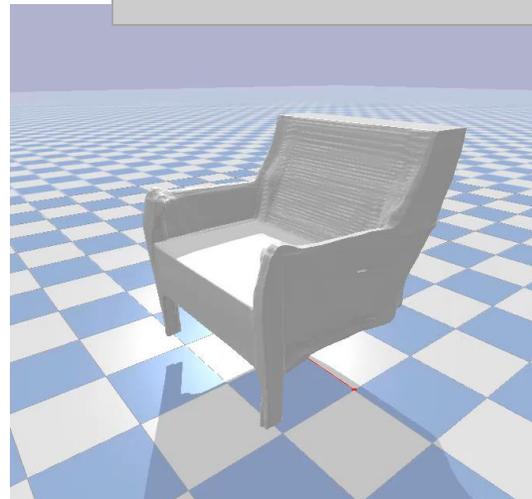
- + **Easy-to-use and mature**
- **Non-differentiable:** need to be combined with gradient approximation methods (instability of numerical gradients).



Non differentiable simulator (e.g., PyBullet)

Our contribution:

- Build a **differentiable** point-based physical simulator
- Learn generative network DeepSDF [1] with **online physical simulation**.



Differentiable simulator Ψ

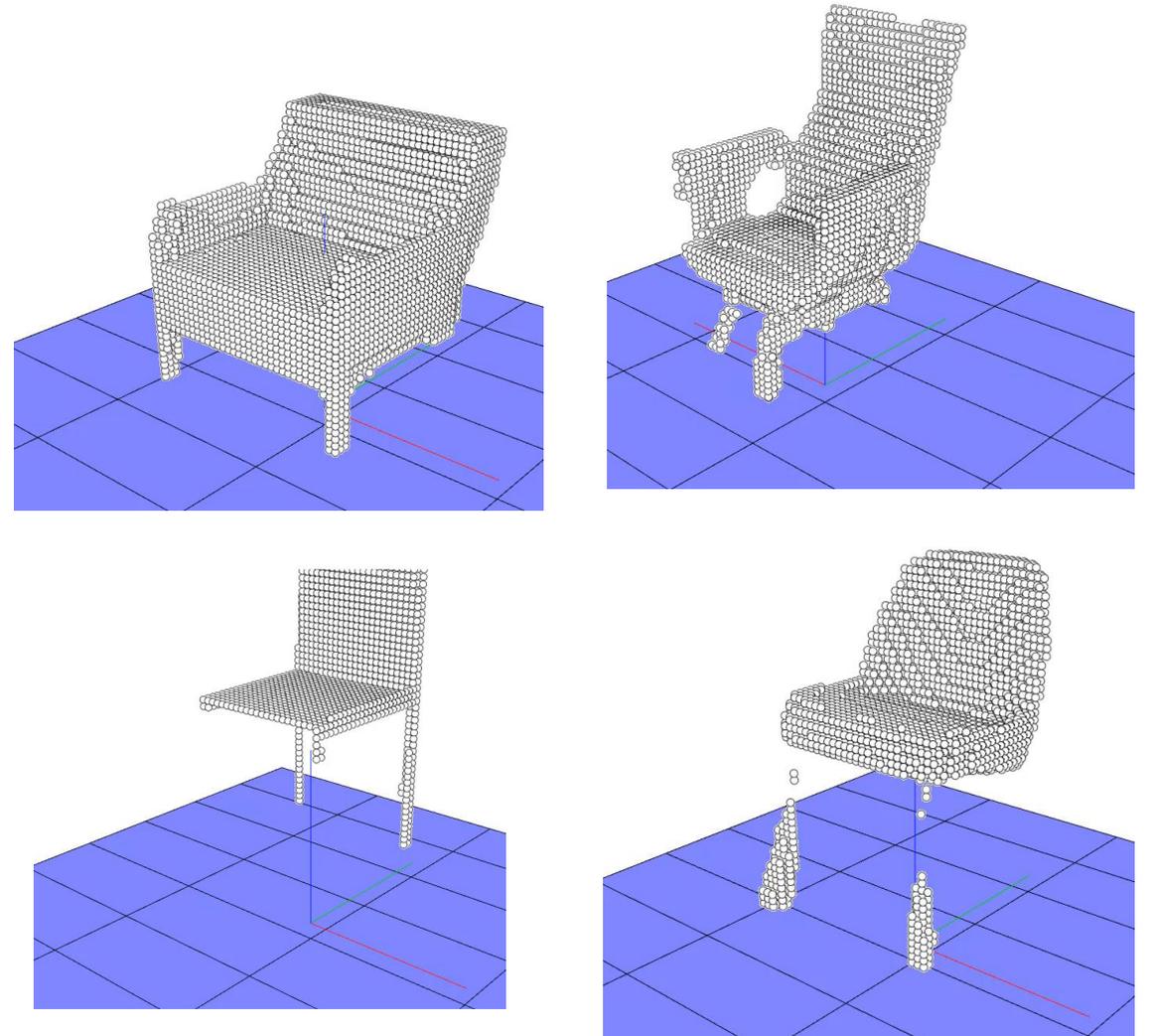
We build a **differentiable simulator Ψ** using the **DiffTaichi [1]** framework :

$$\Psi(\mathcal{C}) = \{(p_t, r_t); t \in [1, T]\}$$

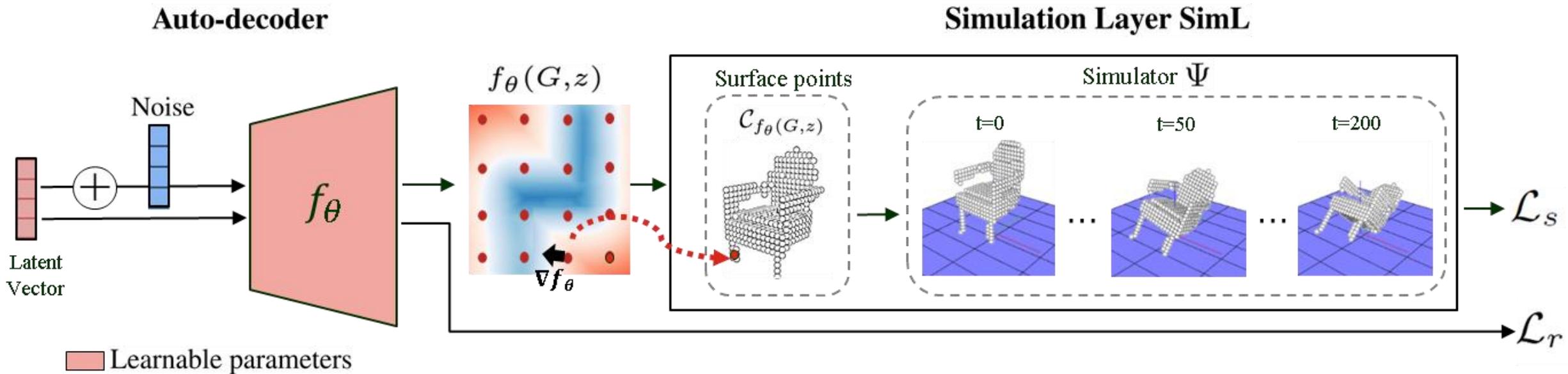
DiffTaichi naturally supports simulation of a **point cloud \mathcal{C}** . We simulate \mathbf{p}_t and \mathbf{r}_t : the *position* and the *rotation* of \mathcal{C} center of mass during simulation.

[1] DiffTaichi: Differentiable Programming for Physical Simulation.
Hu et al., ICLR, 2020

M. Mezghanni, et al. "Physical simulation layer for accurate 3d modeling," CVPR 2022.



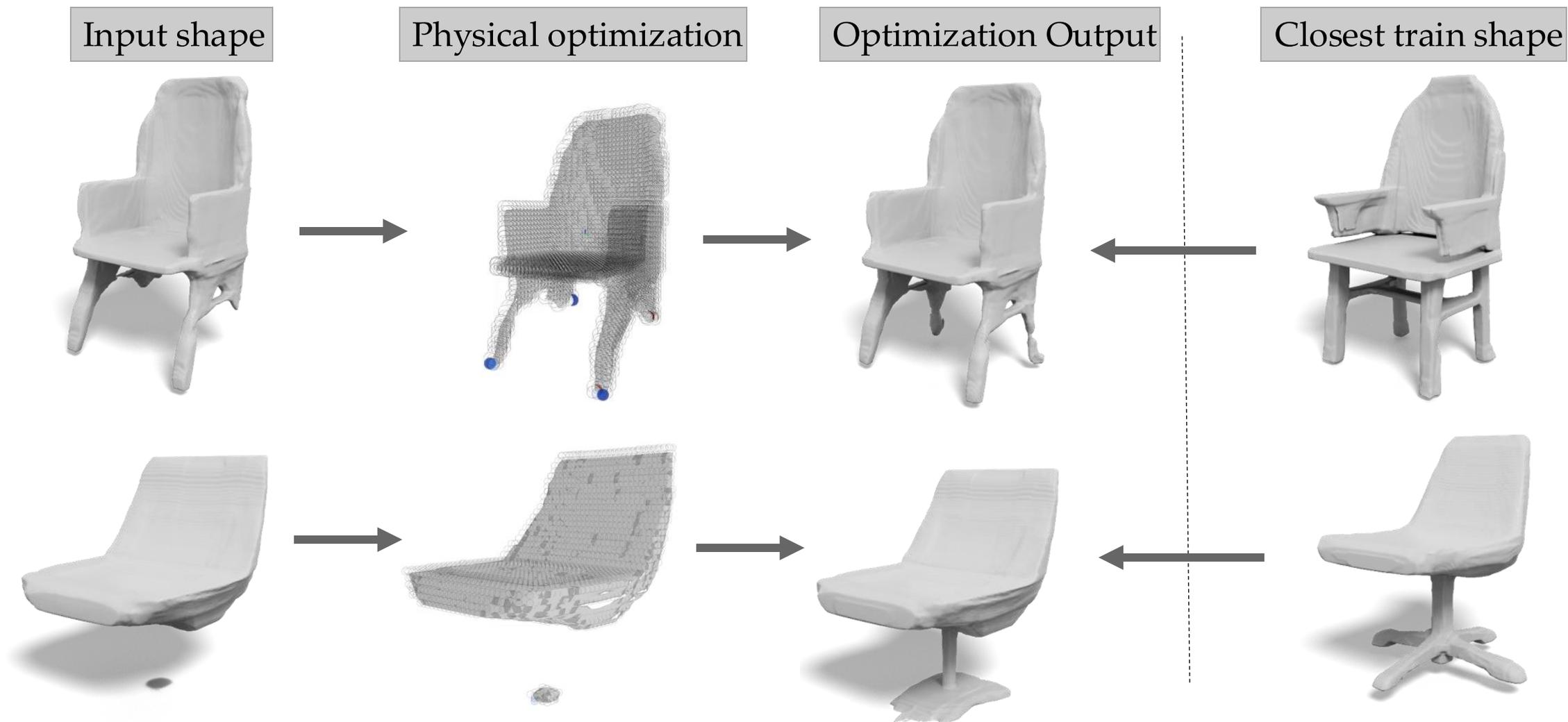
Learning Framework



We train the auto-decoder by jointly optimizing a reconstruction and stability-based losses:

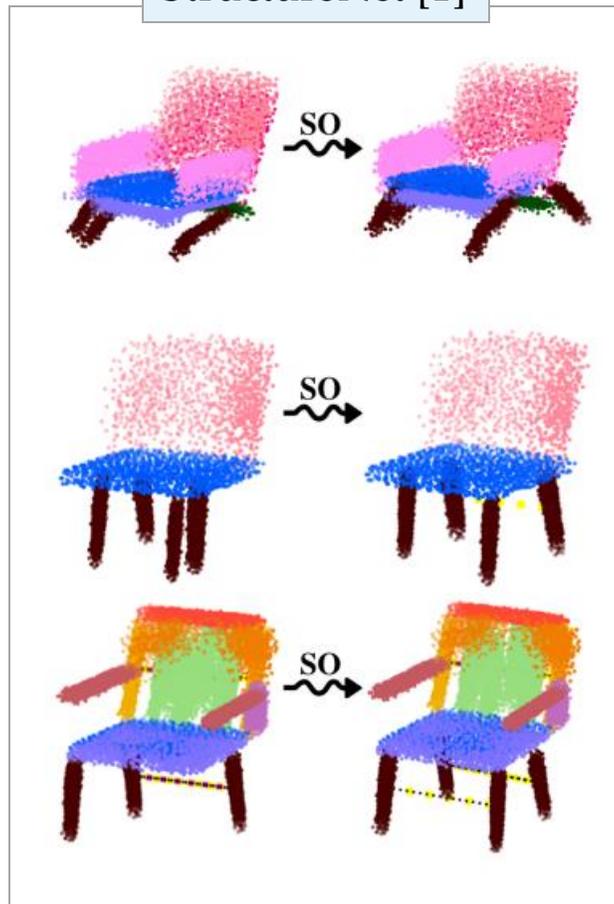
$$\mathcal{L} = \mathcal{L}_r + \alpha_s \mathcal{L}_s$$

Results: Shape Optimization

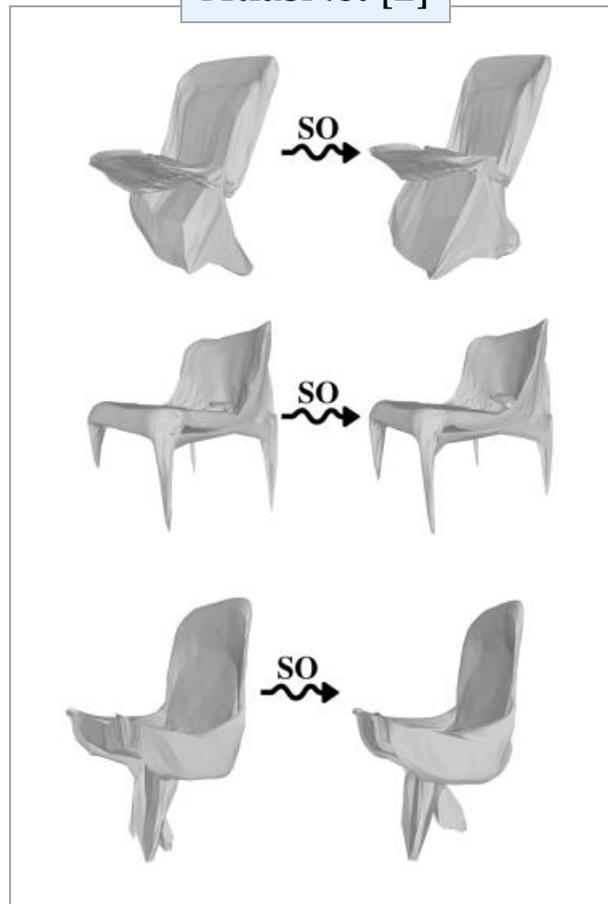


Shape Optimization

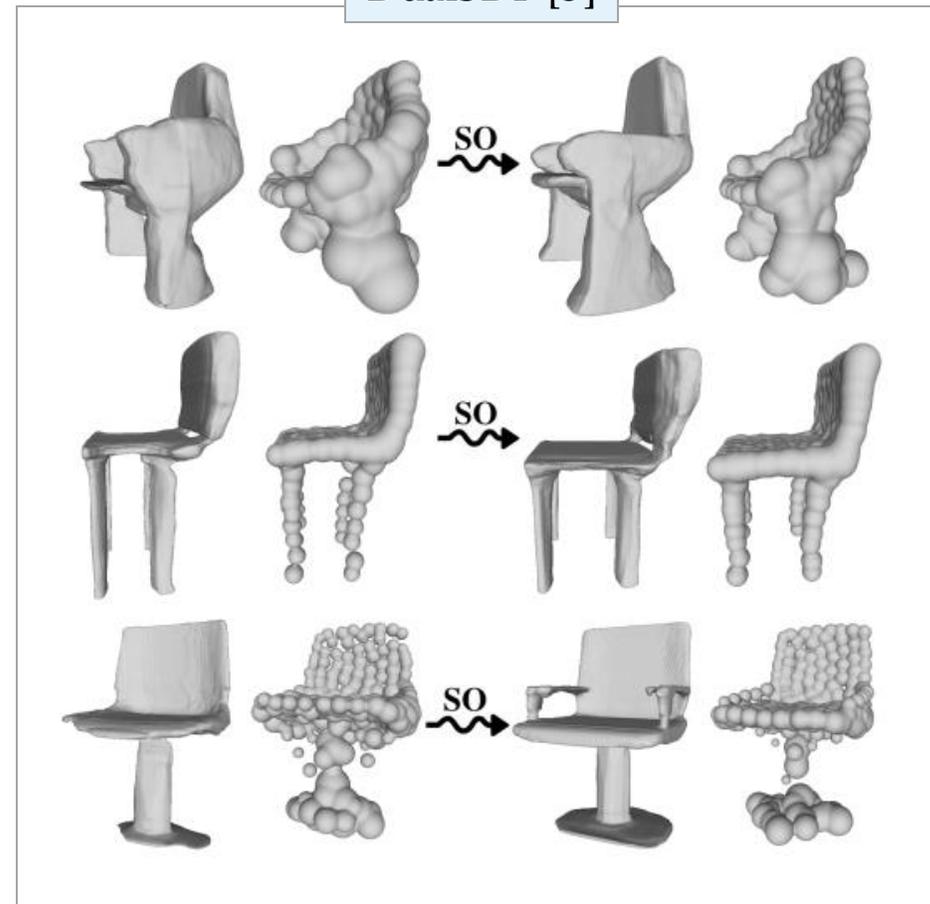
StructureNet [1]



AtlasNet [2]



DualSDF [3]



[1] StructureNet: Hierarchical graph networks for 3d shape generation. Mo et al., SIGGRAPH Asia, 2019

[2] AtlasNet: A Papier-Mache approach to Learning 3D Surface Generation. Groueix et al., CVPR, 2018

[3] Dualsdf: Semantic shape manipulation using a two-level representation. Hao et al., CVPR, 2020

Shape Reconstruction





DeBaRA: Denoising-Based 3D Room Arrangement Generation

Léopold Maillard^{1,2}, Nicolas Sereyjol-Garros, Tom Durand², Maks Ovsjanikov¹

¹LIX, École Polytechnique, IP Paris ²Dassault Systèmes



NeurIPS 2024

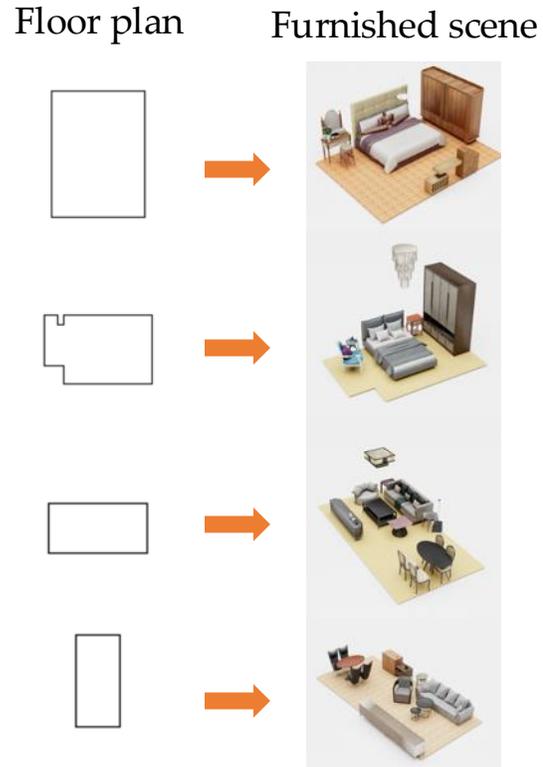


INSTITUT
POLYTECHNIQUE
DE PARIS

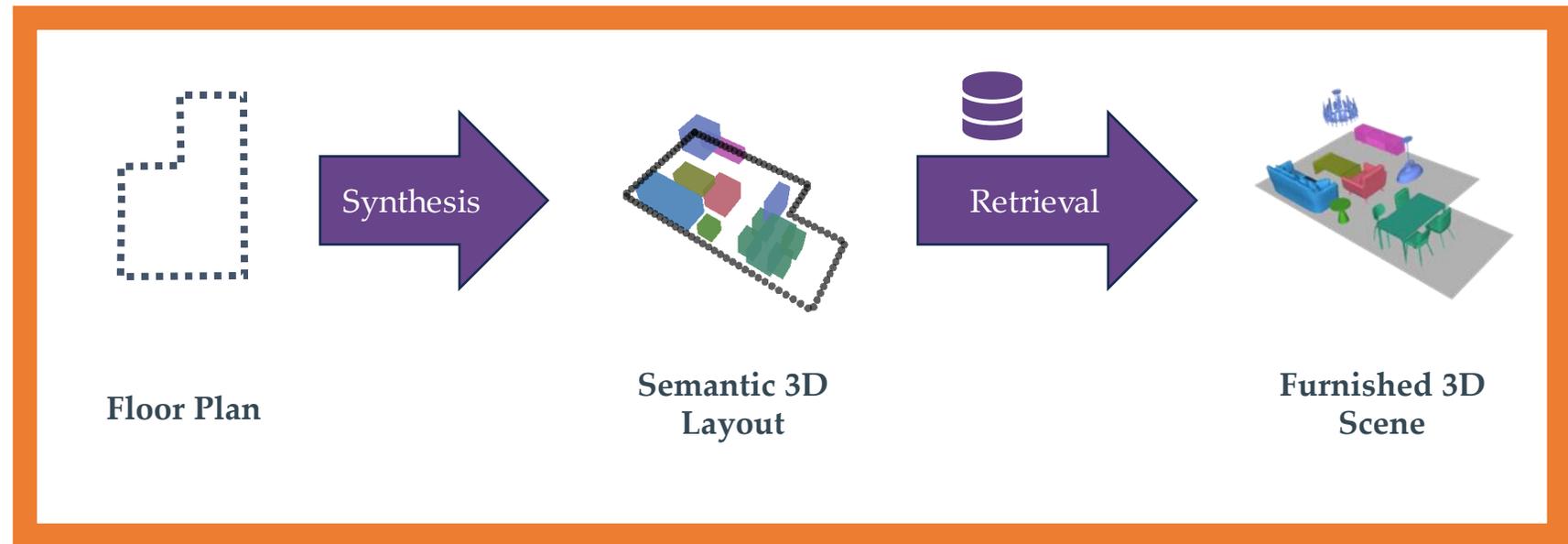


Task: Controllable 3D Indoor Scene Synthesis

Goal



Basic Pipeline



Motivation: Controllable 3D Indoor Scene Synthesis

Challenges

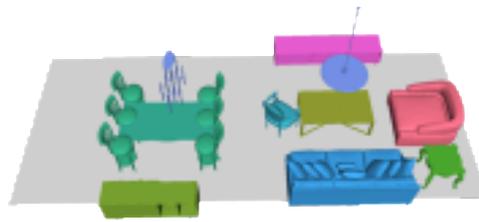
- Inherent complexity of object **interactions**.
- Requirement to fulfill **spatial**, **ergonomic** and **functional** constraints.
- Limited amount of **training data**.

Background

- Existing methods are either **autoregressive** or use **diffusion models** for all object attributes jointly

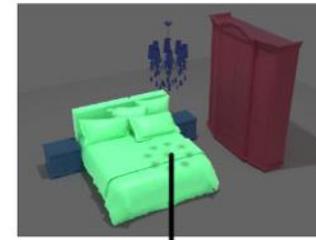


Autoregressive



Off-the-shelf Diffusion

Object Parametrization



Location	$\mathbf{l}_i \in \mathbb{R}^3$
Size	$\mathbf{s}_i \in \mathbb{R}^3$
Orientation	$\theta_i \in \mathbb{R}$
Class	$\mathbf{c}_i \in \mathbb{R}^C$
Shape code	$\mathbf{f}_i \in \mathbb{R}^F$

Object Feature

$$\mathbf{o}_i = [\mathbf{l}_i, \mathbf{s}_i, \cos \theta_i, \sin \theta_i, \mathbf{c}_i, \mathbf{f}_i] \in \mathbb{R}^D$$

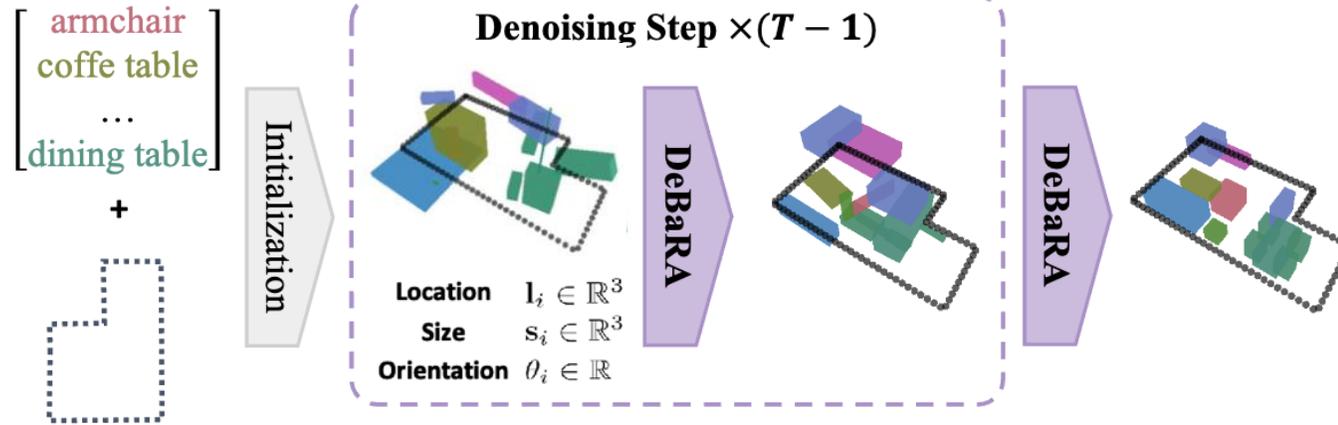
Denoising in a high-dimensional space



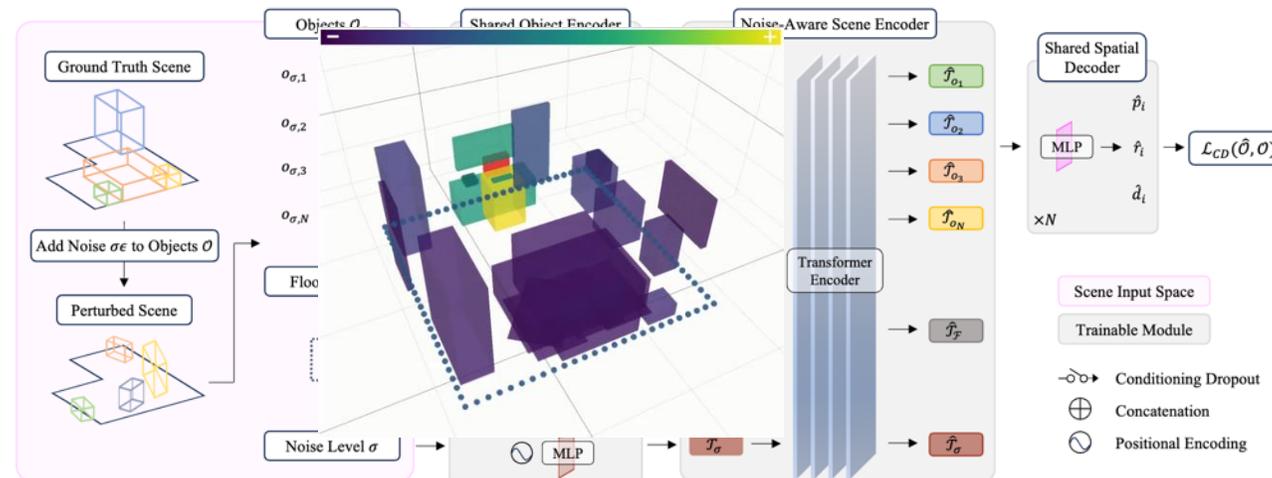
Mixing spatial and semantic features

Our Approach: Separating Geometry and Semantics

Key idea:
 Only denoise the spatial features. Treat the semantic features (object categories) as conditioning.



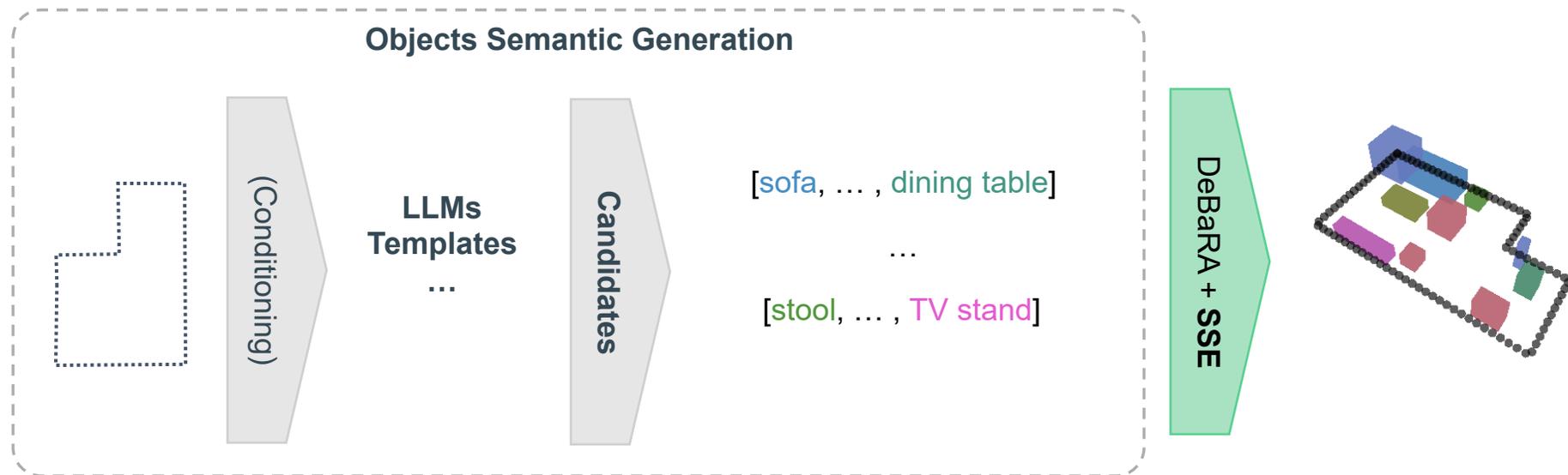
Architecture



How to obtain the conditioning signal?

Input set of object categories can be *provided* by external sources such as a LLM [3].

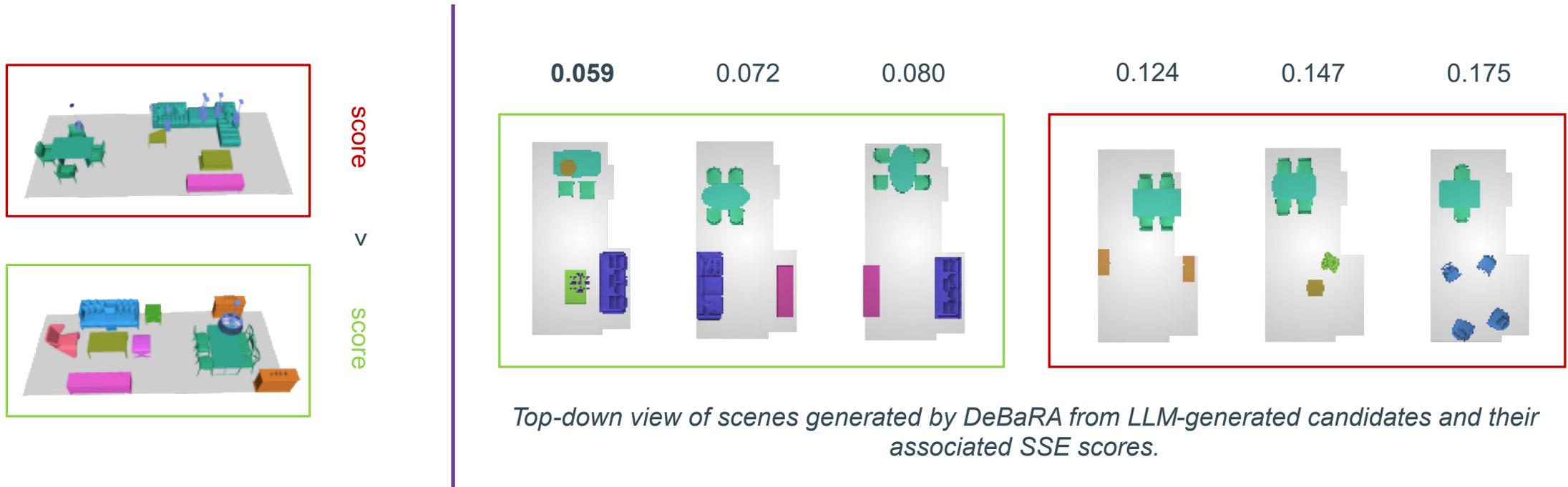
Alternatively, we propose a Self Score Evaluation (SSE) to select the sets that lead to the most realistic scenes. SSE uses density estimation with a trained model.



[3] Feng et al. *LayoutGPT: Compositional Visual Planning and Generation with Large Language Models*, in NeurIPS 2023

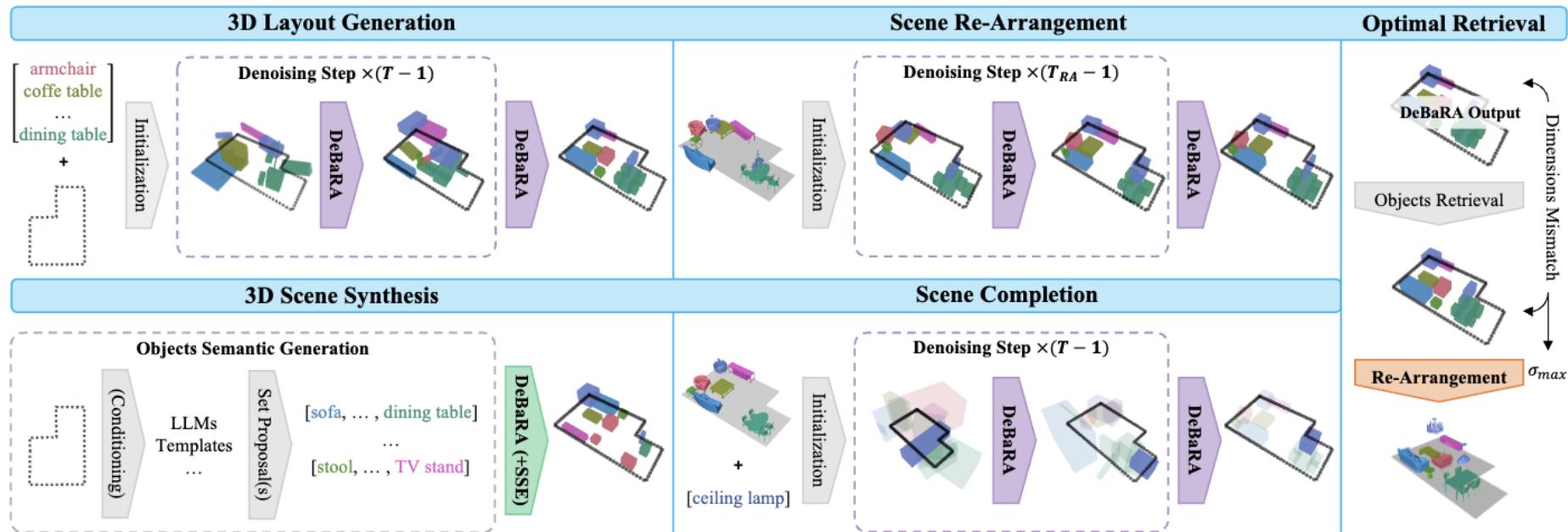
How to obtain the conditioning signal?

Candidate sets of object categories can be automatically generated by a LLM, and using SSE, further **selected** to generate a plausible 3D layout, or automatically **discarded**.



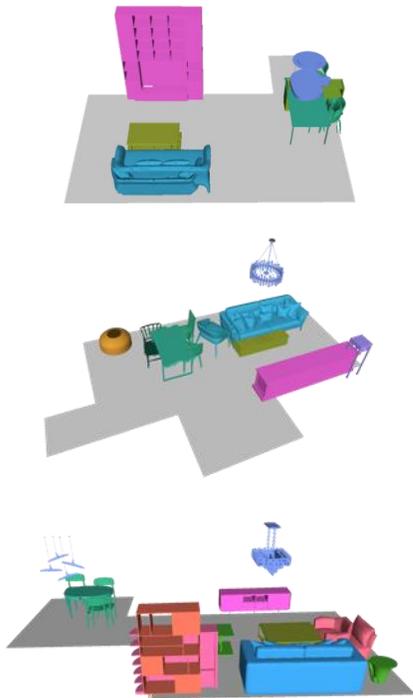
Many Possible Applications

A single pre-trained model can be used for several downstream applications.

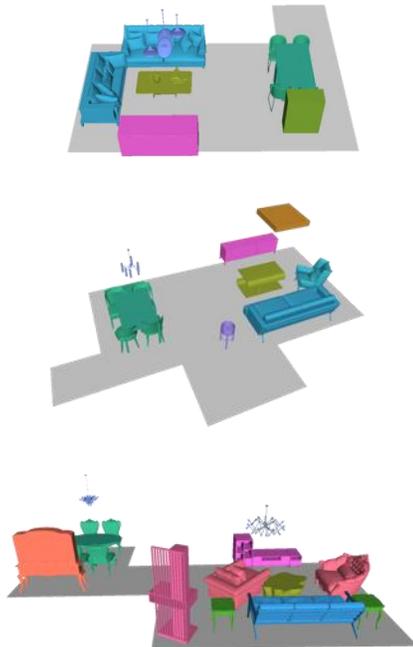


Results – 3D Layout Generation, Synthesis, and Re-arrangement

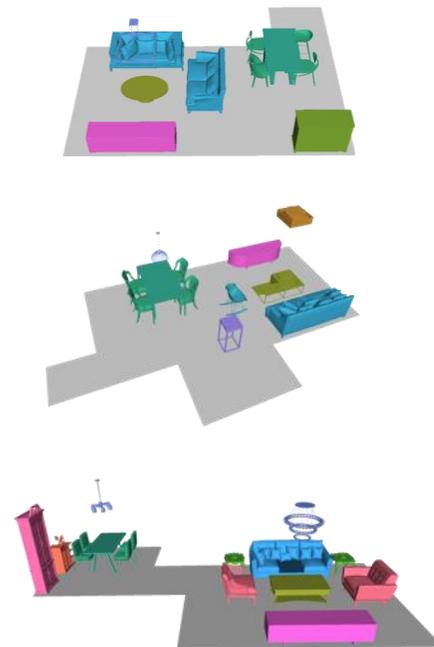
Improved Accuracy in 3D Layout Generation, Scene Synthesis, and Re-arrangement.



ATISS



DiffuScene



DeBaRA

Methods	Living Rooms			
	FID (↓)	KID (↓)	SCA (%)	OBA (↓)
LayoutGPT [9]	35.53	13.69	72.8	2913.6
ATISS [38]	25.67	8.91	71.8	857.3
DiffuScene [51]	21.54	6.40	69.7	341.1
DeBaRA (ours)	18.89	3.57	68.3	167.8

Methods	Dining Rooms			
	FID (↓)	KID (↓)	SCA (%)	OBA (↓)
LayoutGPT [9]	32.80	8.99	67.6	2447.4
ATISS [38]	28.05	9.26	63.2	702.4
DiffuScene [51]	23.06	5.35	57.7	266.4
DeBaRA (ours)	22.04	4.41	52.4	132.8

Especially strong improvement in physical consistency.

Thank You

Questions?

Acknowledgements:

M. Mezghanni, L. Maillard, M. Boulkenafed, ...

Work supported by the ERC Starting Grant StG-2017-758800 (EXPROTEA),
ERC Consolidator VEGA and the ANR AI Chair AIGRETTE.

