A Survey on Geometry Deep Learning PPT Notes

- 2D Images deep learning
- 3D shapes deep learning

- Unlike 2D images, which can be uniformly represented by a regular grid of pixels, 3D shapes have **various representations**. Methods to 3D shapes deep learning vary in different representations.

Research Directions of GDL

- 3D shape classification
- 3D shape recognition
- 2D to 3D shape reconstruction
- 3D shape Generation
- 3D shape completion
- Point cloud & 3D mesh segmentation
- Pose estimation
- …

Representation

- Voxels
- Depth and multi-View Images
  - e.g. RGB-D
- Surfaces
  - e.g. Mesh, point cloud
- Implicit representation
- Structured representation
- Deformation-based representation

Representations

3D Data type	Advantages	Disadvantages
Voxel	i. An extension of the concept of pixel ii. Defined on a fixed regular grid. iii. Defined on a fixed regular grid	i. Computational expensive ii. Memory management problem. iii. Low resolution output
Point cloud	i. Sampled 3D point coordinates to represent surfaces ii.Easily generated by scanners	i.It is irregular in nature ii.Difficult to process iii.Sparse representations of shapes
SDF	i. Function that assigns a 3D point to a real value ii.It possesses unlimited resolution iii. Provide easy way to deal with topology changes iv. It can be used to model arbitrary shape topology v. It ensures watertight surface vi. Easy integration with learning framework vii. SDF preserve the surface information viii. Ease of conversion to other 3D data representations	i. Suffers aliasing errors under low resolution ii.Memory inefficient under high resolution iii.Unclear on how it represent a weighted mesh
Mesh	i. Uses vertices,edges and faces to defines object ii. Higher quality 3D shapes iii. Less memory and computational cost compared to voxel and point cloud	i. Irregular in nature ii. Not uniquely defined iii. Integration with learning frameworks not easy iv. Require shape deforming, hence it does not allow arbitrary topologies

Fig. 0: Comparisons of different 3D shape representations. [*]

Dense Voxel

- [1] 3D ShapeNets. Wu et al. CVPR. 2015.
  - Use a Convolutional Deep Belief Network.
  - Input depth maps → convert to voxel representation→ Output category labels and predicted 3D shape.

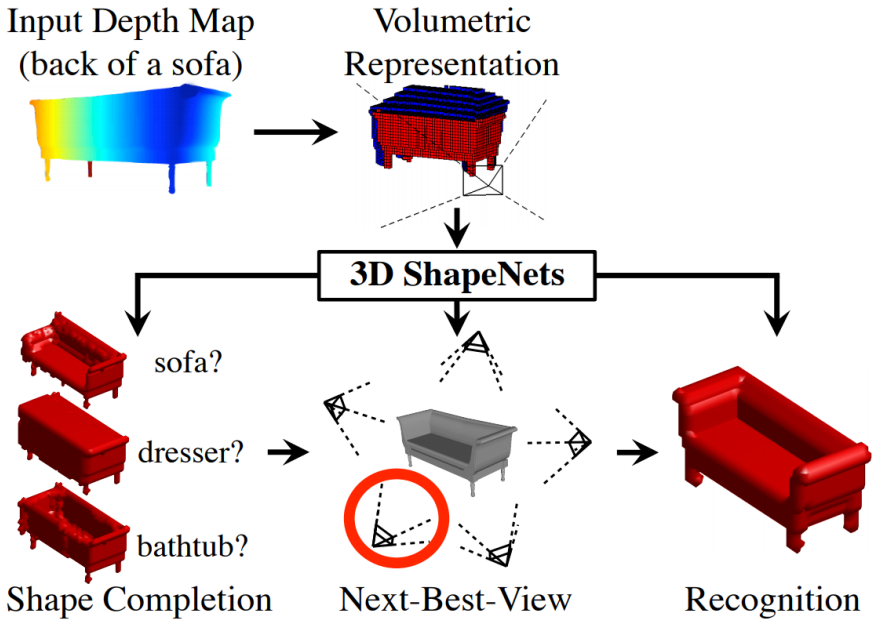


Fig. 1: Usages of 3D ShapeNets. Given a depth map of an object, we convert it into a volumetric representation and identify the observed surface, free space and occluded space. 3D ShapeNets can recognize object category, complete full 3D shape, and predict the next best view if the initial recognition is uncertain. Finally, 3D ShapeNets can integrate new views to recognize object jointly with all views.

http://3dshapenets.cs.princeton.edu/

Dense Voxel

- [2] VoxNet. Maturana et al. IROS. 2015
- [3] VConv-DAE. Sharma et al. 2016
  - An autoencoder model.
  - Without labels. **Unsupervised**.
- [4] 3D R2N2. Choy et al. ECCV. 2016
  - Reconstruction: Input images→Output occupancy grid.
  - 2D Image encoder→3D-LSTM→3D Deconvolutional Decoder
- [5] 3D-GAN & 3D-VAE-GAN. Wu et. al
  - Reconstruction

Dense Voxel

The main **challenges** affecting performance include overfitting, **orientation**, **data sparsity**, and **low resolution,**

e.g. (Image based) *Multi-View CNNs* performs better on classification tasks than naïve volumetric networks.

- [5] Volumetric and multi-view CNNs. Qi, et al. CVPR. 2016.
  - Propose several separate approaches to improve the performance of volumetric and multi-view CNN for 3D object classification.

- [6] ORION. Sedaghat et al. BMVC. 2017
  - Predict class and **orientation**

Sparse Voxel

- Use a sparse, adaptive data structure, the **Octree.**

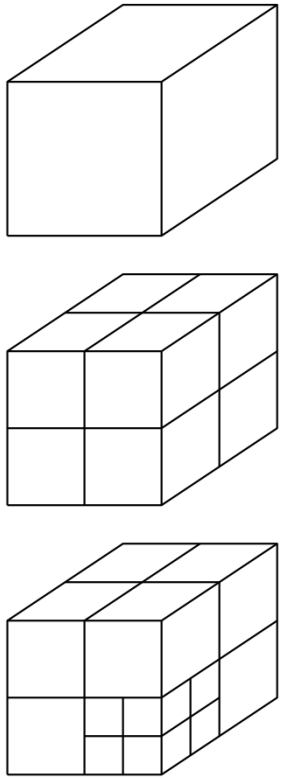


Fig. 2: Visualization of the voxel block octree [7].

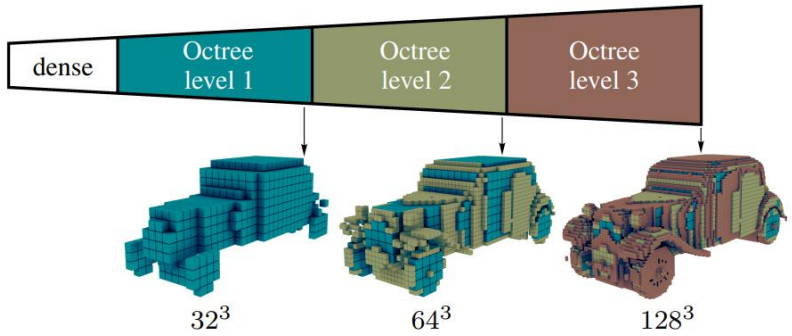


Fig. 3: The proposed OGN represents its volumetric output as an octree. [8] Initially estimated rough low-resolution structure is gradually refined to a desired high resolution. At each level only a sparse set of spatial locations is predicted.

Sparse Voxel

- [7] Hierarchical surface prediction (HSP). (Reconstruction). Hane et al. 2017.
  - High resolution around the surfaces (up to 256^3), coarse interior and exterior voxels.
- [8] Octree Generative Network (OGN). Tatarchenko et al. ICCV. 2017.
  - A deep convolutional decoder architecture that yields an octree as output.
  - 3 categories of nodes: full, empty, mixed. Feature map stored in the forms of hash tables indexed by spatial position and octree levels.
  - *OGN-Conv* is designed to convert convolution operations into matrix multiplication.

Sparse Voxel

- The reconstruction result of OGN [8]

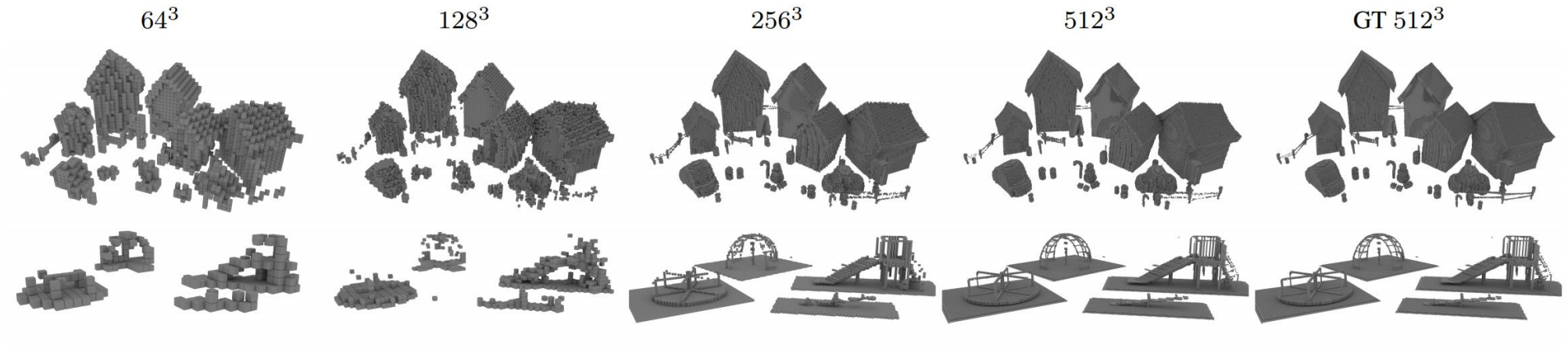


Fig. 4: OGN is used to reproduce large-scale scenes from the dataset.

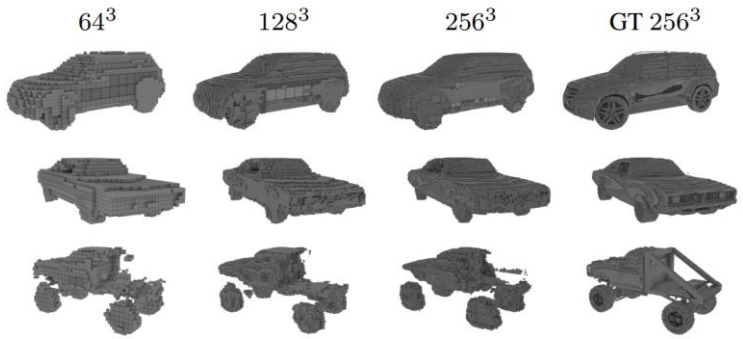


Fig. 5: Samples from the ShapeNet-cars dataset and the result from OGN

Sparse Voxel

Efforts were made to design new structures for octrees and special operations on octrees.

- [9] OctNet. Riegler et al. CVPR. 2017.
  - Shallow octree in cells of a regular 3D grid.
- [10] Octree-based CNNs (O-CNN). Wang et al. TOG. 2017.
  - Removes pointers like a shallow octree and stores the octree data and structure using a series of vectors

Point-based

- Typically point cloud or point set. Can be more **easily obtained** from 3D scanners than other 3D representations.

- Due to its **unordered** and **irregular** structure, it is difficult to cope with using traditional deep learning method.

Point-based

- [11] PointNet. Charles et al. CVPR. 2017.
  - Used for **shape classification & segmentation**. The first successful deep network to directly process point clouds without unnecessary rendering.
  - See also PointNet++ [12]. NIPS. 2017. A hierarchical structure is introduced allowing it to capture features at different scales.

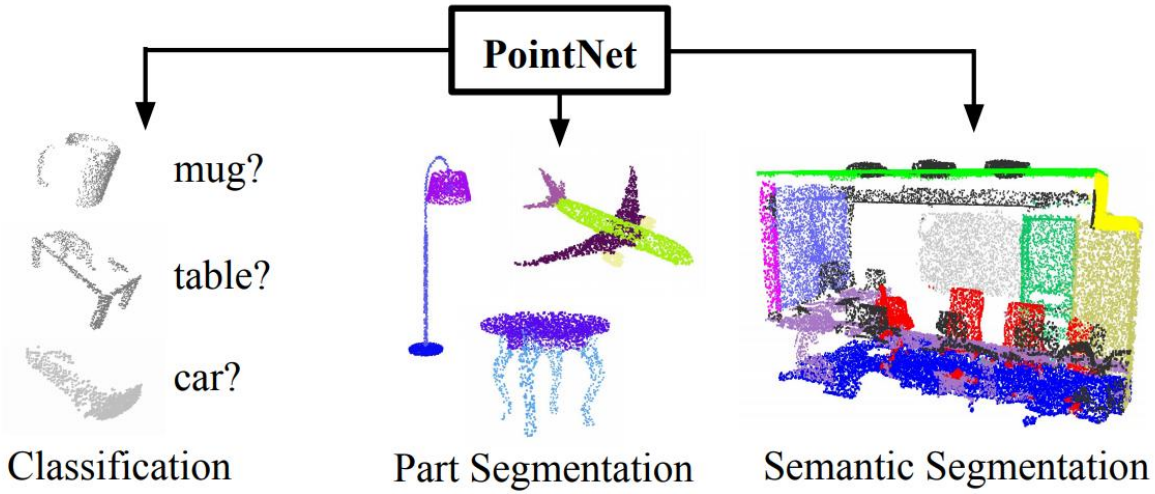


Fig. 6: Applications of PointNet.

Point-based

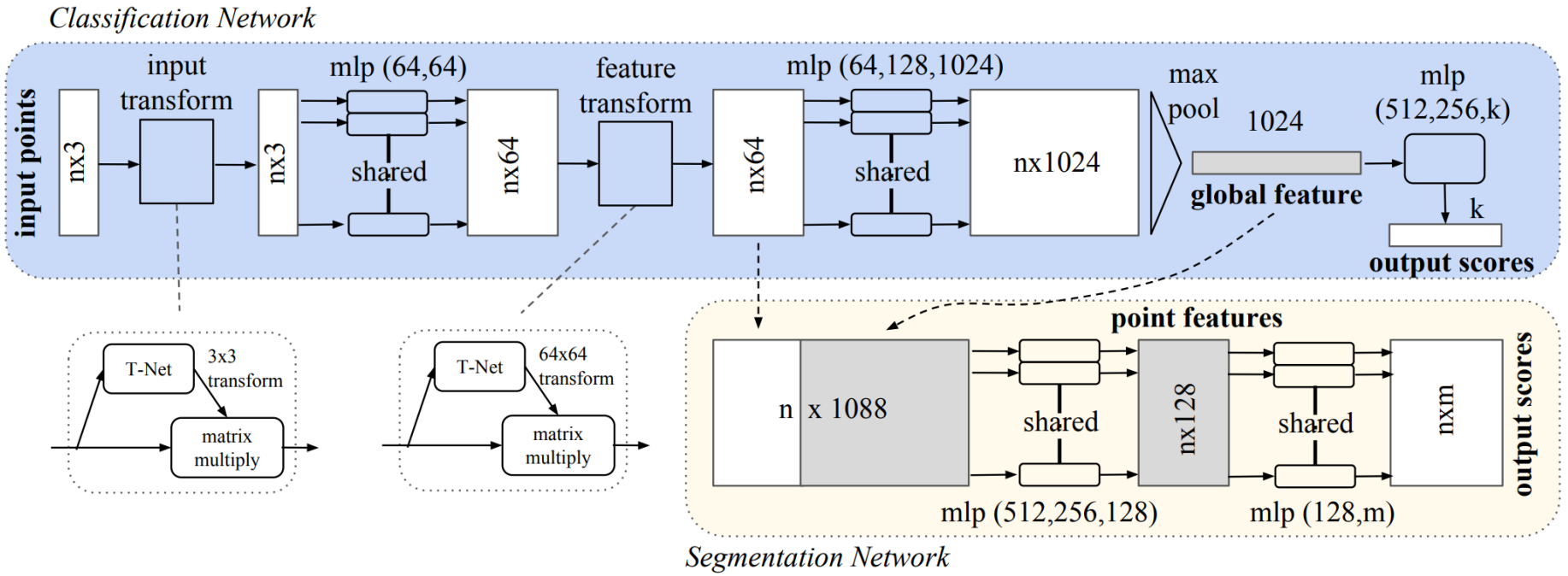


Figure 7. PointNet Architecture. [11] The classification network takes n points as input, applies input and feature transformations, and then aggregates point features by max pooling. The output is classification scores for k classes. The segmentation network is an extension to the classification net. It concatenates global and local features and outputs per point scores. "mlp" stands for multi-layer perceptron, numbers in bracket are layer sizes. Batchnorm is used for all layers with ReLU. Dropout layers are used for the last mlp in classification net.

Point-based

Some focus on apply **CNNs** to point clouds.

- [13] PointCNN. Li et al. NIPS. 2018.
  - Designed $\mathcal{X}$-transformation. Each feature matrix is multiplied by the $\mathcal{X}$-transformation matrix before passing through the convolutional operator, **guaranteeing equivariance for different point orders**.

- [14] DGCNN. Wang et al. TOG. 2018.
  - A dynamic graph CNN architecture
  - First connects neighboring points in spatial or semantic space to generate a graph, and then captures local geometric features by applying the EdgeConv operator to it.

Point-based

Other NNs approaches

- [15] Kd-networks. Klokov et al. ICCV. 2019.
  - Based on k-d-trees.

Point-based

- [16] Voxel VAE Net (VV-Net). Meng et al. ICCV. 2019.
  - Uses a latent code computed by an **Radial Basis Function interpolated Variational Auto-Encoder (RBF-VAE)** to describe point distribution within a voxel.
  - The output is combined with *PointNet* for better segmentation performance.

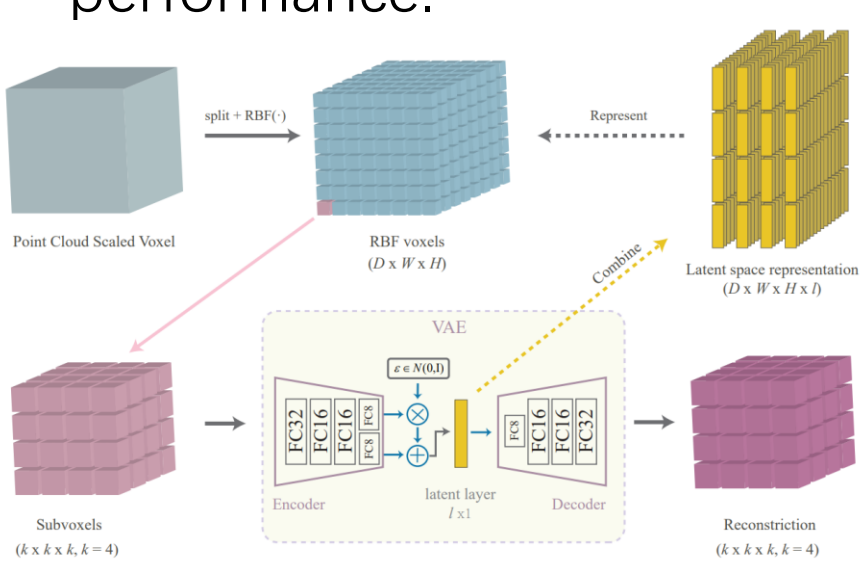
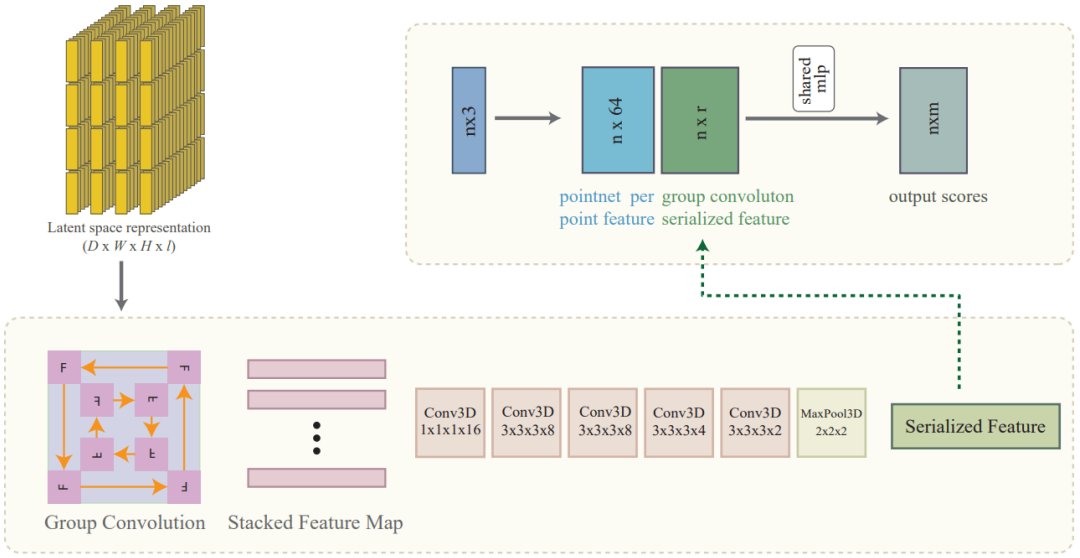


Fig. 8: RBF-VAE Architecture



Segmentation Network Architecture

Mesh-based

- Unlike point-based representations, mesh-based representations provide **connectivity** between neighboring points.
- Learning 3D meshes is a difficult task in CV and CG…

- It has been used in:
  - 3D mesh generation.
  - 3D shape completion,
  - 3D shape segmentation,
  - 3D action recognition,
  - scene generation,
  - human-object interaction,
  - 3D scene augmentation,
  - 3D human pose estimation,
  - …

Mesh-based

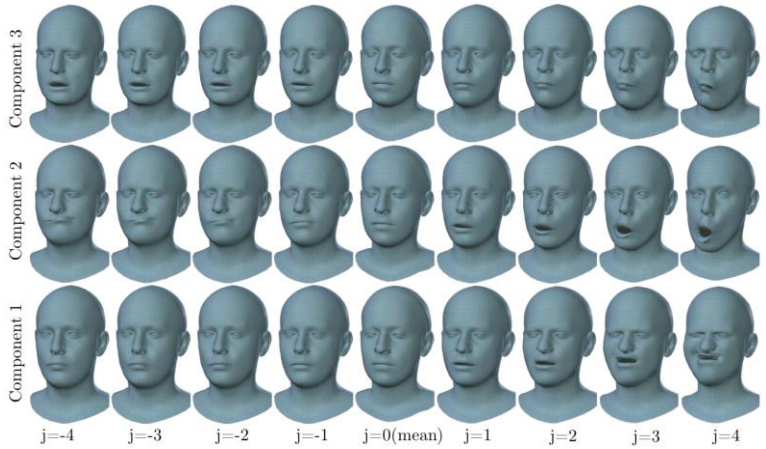


Fig. 9: 3D face generation. [*]

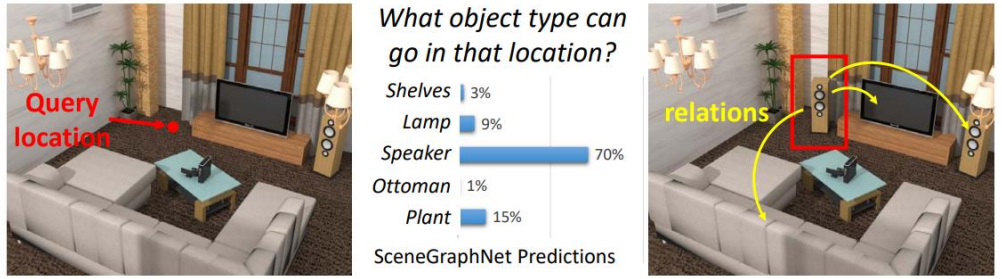


Fig. 10: SceneGraphNet for scene augmentation. [*]

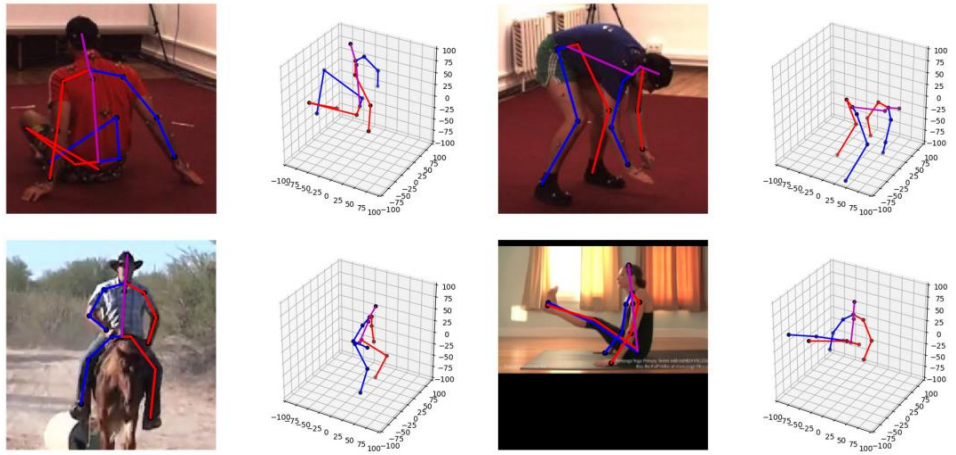


Fig. 11: Semantic GCNs for 3D Human Pose Regression. [*]

Mesh-based

Directly applying CNNs to irregular mesh data structure is non-trivial. A handful approach is based on **parametric representations** for meshes, mapping 3D shape surfaces to 2D domains.

- [17] SurfNet. Sinha et al. CVPR. 2017. (Reconstruction)
  - 2D image to 3D shape surfaces reconstruction using deep residual network.
- [18] DeepPano. Shi et al. SPL. 2015. (Recognition)
  - Project 3D models into cylinder panoramic images, which are then processed by CNNs.

Mesh-based

- [19] TextureNet. Huang et al. CVPR. 2019.
  - Takes as input a 3D textured mesh, extract features from high-resolution signals (e.g. colored map) associated with 3D surface meshes. Outputs are learned features for a dense set of sample points that can be used for semantic segmentation and other tasks.

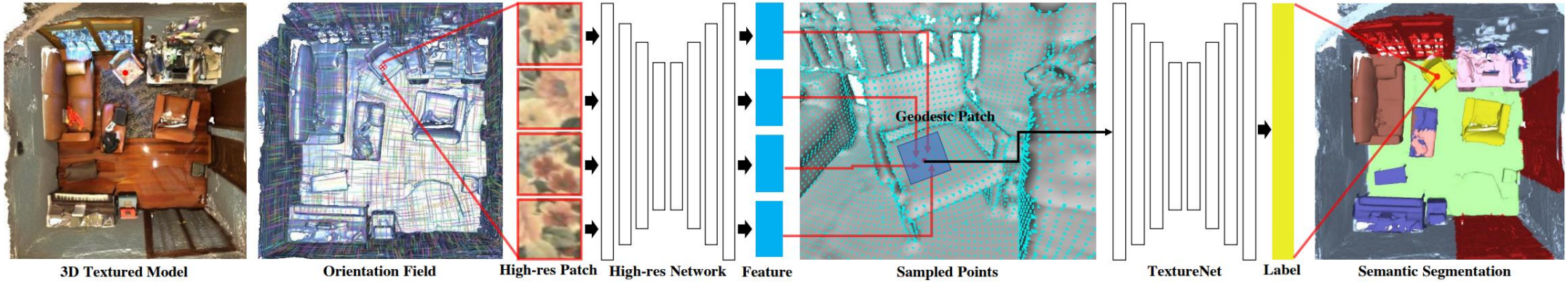


Fig. 12: TextureNet. The mesh is **parameterized** with a consistent 4-way rotationally symmetric (4-RoSy) field, which is used to extract oriented patches from the texture at a set of sample points. Networks of 4-RoSy convolutional operators extract features from the patches and used for 3D semantic segmentation.

Mesh-based

Mesh structure is constructed with vertices and edges, and can be seen as **graph**. Some models have been proposed based on the **graph spectral theorem**. They generalize CNNs on graphs by eigen-decomposition of Laplacian matrices, generalizing convolutional operators to the spectral domain of graphs.

- [20] FeaStNet. Verma et al. CVPR. 2018.
  - A novel **graph-convolution operator** is proposed to establish correspondences between filter weights and graph neighborhoods with arbitrary connectivity.

Mesh-based

- [21] MeshCNN. Hanocka et al. TOG. 2019.
  - Unlike other graph-based methods, it focuses on processing features stored in edges, using a **convolution operator applied to the edges** with a fixed number of neighbors and **a pooling operator based on edge collapse**.

Mesh-based

Several works have been designed using **2-manifolds** with a series of refined CNN operators adapted to such nonEuclidean spaces.

- [22] Geodesic CNN (GCNN). Masci et al. ICCV workshop. 2015.
  - Extract and discretize local geodesic patches and apply convolutional filters to these patches in polar coordinates.

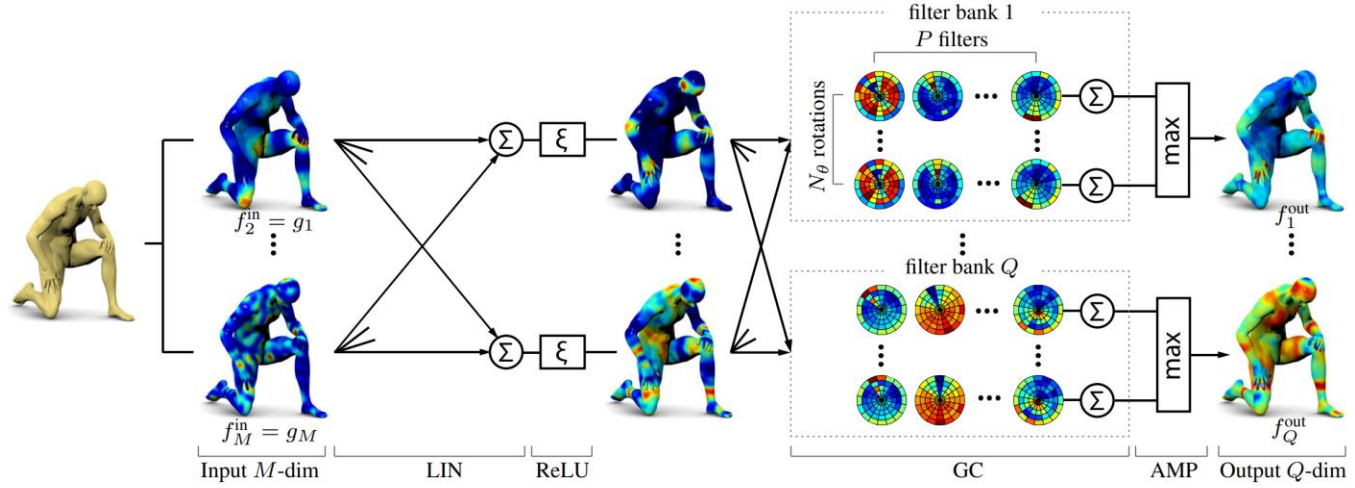


Figure 3: The simple GCNN1 architecture containing one convolutional layer applied to $M = 150$-dimensional geometry vectors (input layer) of a human shape, to produce a $Q = 16$-dimensional feature descriptor (output layer).

Mesh-based

- Localized spectral CNNs. Boscaini et al.
- Anisotropic convolutional neural networks (ACNNs). Boscaini et al.
- directionally convolutional networks (DCNs). Xu et al.
- MoNet. Moti et al.
- SplineCNN
- Laplacian pooling network (LaplacianNet).

Mesh-based

Generative models for mesh-based representation.

- [23] Pixel2Mesh. Wang et al. ECCV. 2018.
  - Based on a graph-based convolutional networks (GCNs).
  - Reconstruct 3D shapes from single images. It generates the target triangular mesh by **deforming an ellipsoidal template.**
  - See also [24] Pixel2Mesh++ by Wen et al (ICCV 2019). which is extended to reconstruction from multi-view images.

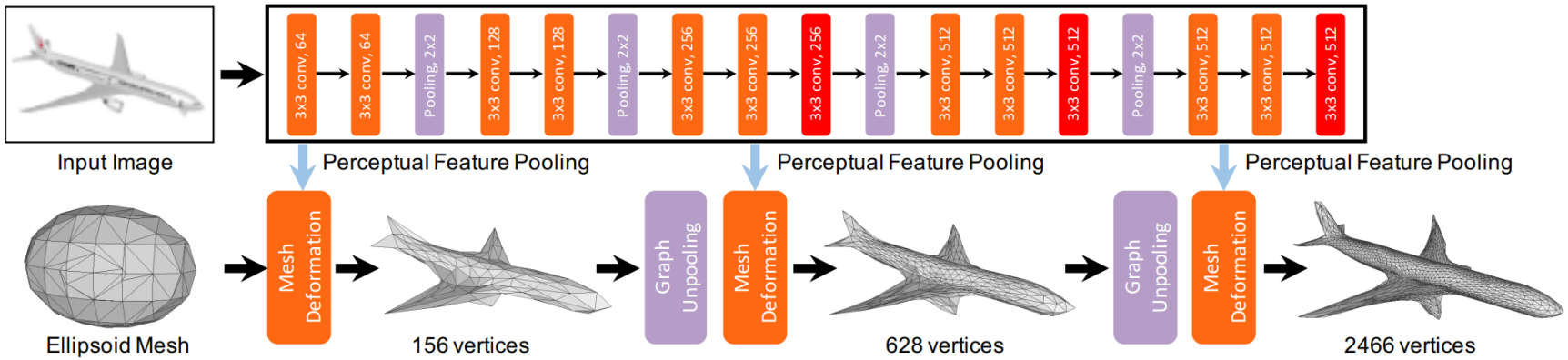


Fig. 14: The cascaded mesh deformation network.

Mesh-based

- [25] AtlasNet. Groueix et al. CVPR. 2018.
  - Generates 3D surfaces from multiple patches.
  - Learns to convert 2D square patches into 2-manifolds to cover the surface of 3D shapes using an MLP (multi-layer perceptron)

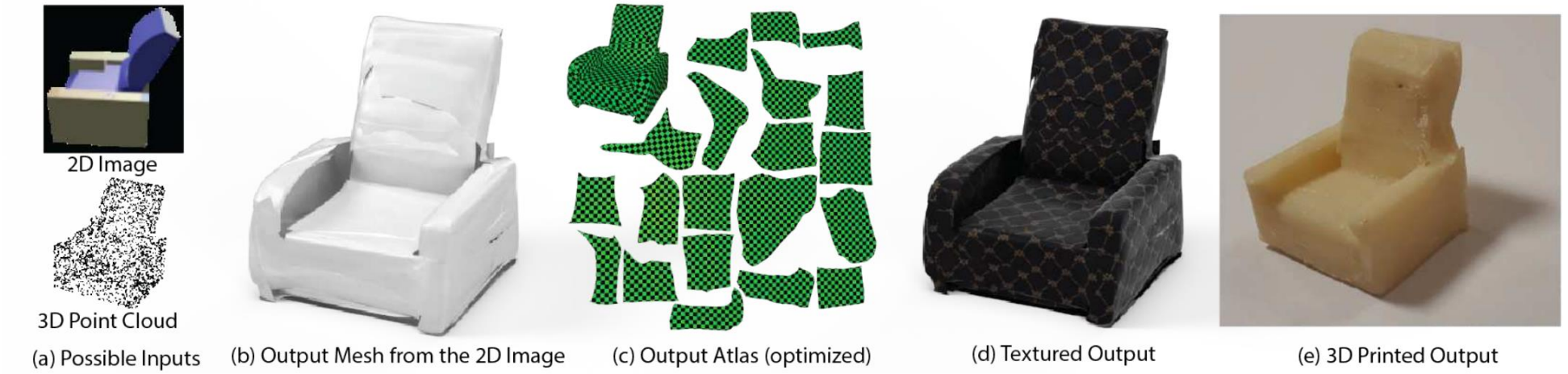
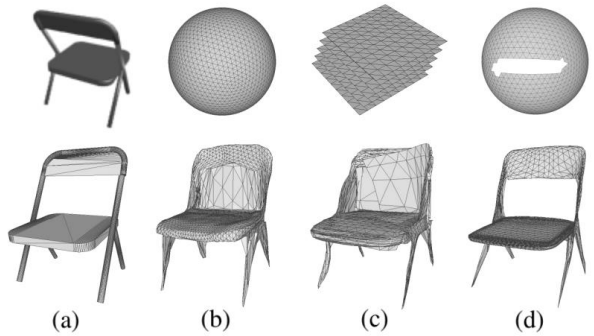


Fig. 15: Given input as either a 2D image or a 3D point cloud (a), we automatically generate a corresponding 3D mesh (b) and its atlas parameterization (c). We can use the recovered mesh and atlas to apply texture to the output shape (d) as well as 3D print the results (e).

Mesh-based



Methods based on deforming a template cannot well capture **complex topology**. (See in Fig.)

- [26] Deep Mesh Reconstruction from Single RGB Images via Topology Modification Networks. Pan et al. ICCV 2019.
  - Single-view reconstruction method which combines a **deformation network** and a **topology modification network** to model meshes with **complex topology**.

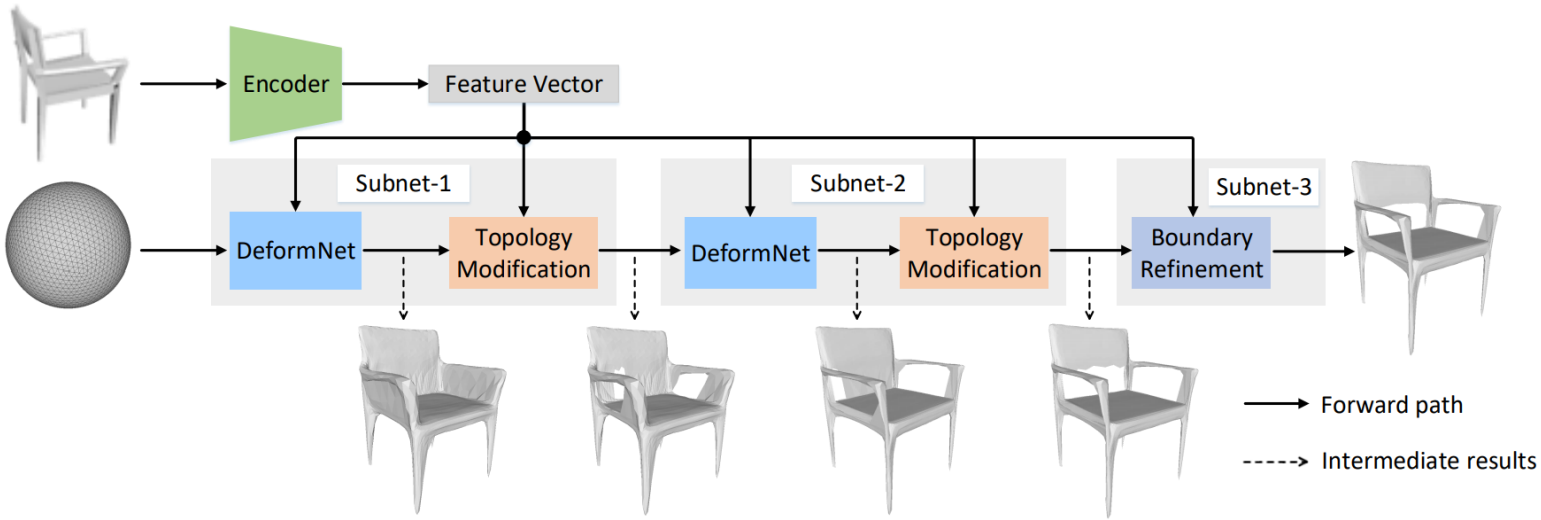


Fig. 16: The overview of Topology-adaptive Mesh Reconstruction pipeline

Mesh-based

- [27] A skeleton-bridged deep learning approach for generating meshes of complex topologies from single RGB images. Tang et al. CVPR. 2019.
  - generating complex topology meshes using a skeleton-bridged learning method, as a skeleton can well preserve topology information.

- [28] PolyGen. Nash et al. 2020.
  - Instead of generating triangular meshes, PolyGen generates a polygon mesh representation.

Implicit Representations

- In addition to explicit representations such as point clouds and meshes, implicit representations have increased in popularity in recent studies. A major reason is that implicit representations are **not limited to fixed topology or resolution**.
- An increasing number of deep models define their own implicit representations and build on them for various methods of shape analysis and generation.

Implicit Representations

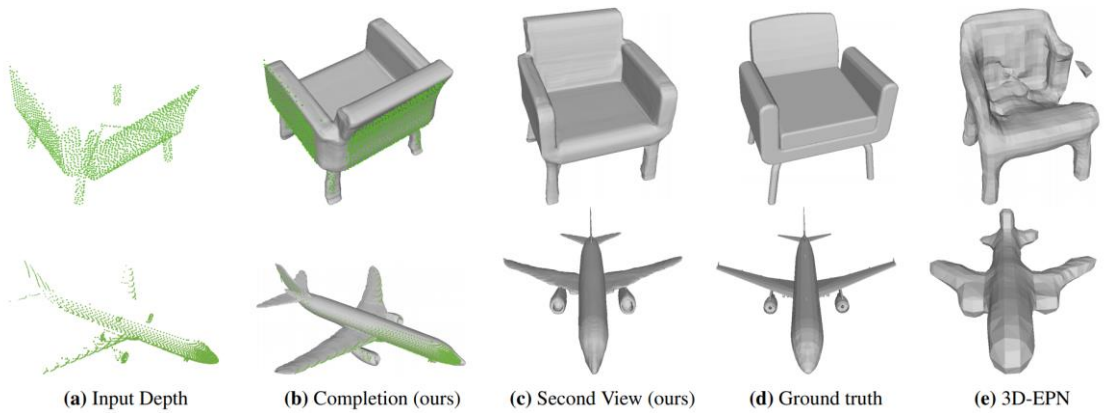
Occupancy and indicator functions are one way to represent 3D shapes implicitly. The occupancy function reflects 3D point status with respect to the 3D shape's surface, where **1** means inside the surface and **0** otherwise.

- [29] Occupancy networks. Mescheder et al. CVPR. 2019.
  - The occupancy network is defined as $f_\theta(p, x) \in [0,1], (p, x) \in \mathbb{R}^3 \times \mathcal{X}$.
  - Then use *Multiresolution IsoSurface Extraction*(MISE) to extract an approximate isosurface $\{p \in \mathbb{R}^3 | f_\theta(p, x) = \tau\}$.

Implicit Representations

Signed distance functions (SDFs) are another form of implicit representation. They map a 3D point to a real value indicating the spatial relation and distance to the 3D surface. For a given 3D point $x \in \mathbb{R}^3$,

$$SDF(x) \begin{cases} < 0, & x \text{ inside the 3D shape} \\ = 0, & x \text{ on the surface} \\ > 0, & x \text{ outside the 3D shape} \end{cases}$$

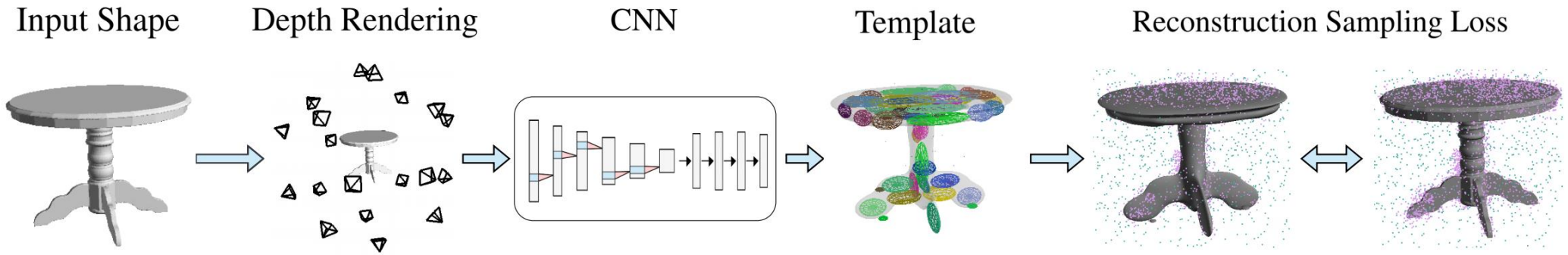


- [30] DeepSDF. Park et al. CVPR. 2019.
  - f_θ is the **auto-decoder** model, a function of a shape latent code z and a query 3D location x, and outputs the shape's approximate SDF $f_\theta(z, x)$.

Implicit Representations

Function sets.

- [31] Structured Implicit Functions (SIFs). Genova et al. ICCV. 2019.
  - Each element is represented by a *scaled axis-aligned anisotropic 3D Gaussian*, and the sum of these shape elements represents the whole 3D shape. The Gaussians' parameters are learned by the CNN.



- [32] Deep SIFs. Genova et al. 2019.
  - SIF to depict coarse shape.
  - *Deep implicit function* (DIF) for local geometry detail.

Deformation-based representations

- Most methods mentioned above focus on rigid 3D models, and pay less attention to deformation of **non-rigid** models. Unlike other representations, deformation-based representations **parameterize the deformation information** and achieve better performance for non-rigid 3D shapes such as articulated models.

Deformation-based

Some **mesh-based** generation methods generate target shapes by deforming a mesh template, and these methods can also be regarded as deformation-based methods.

- [23] Pixel2Mesh. Wang et al. ECCV. 2018.
- [24] Pixel2Mesh++. Wen et al. ICCV. 2019.
- [32] SDM-NET: Deep generative network for structured deformable mesh

Deformation-based

- Efficient and Flexible Deformation Representation for Data-Driven Surface Modeling (RIMD). Gao et al. TOG. 2016.
  - Designed an efficient, rotation-invariant deformation representation called *rotation-invariant mesh difference* (RIMD).

- MeshVAE. Tan et al. CVPR. 2018.
  - Take the RIMD as feature inputs of VAE, and uses fully connected layers for the encoder and decoder.

Deformation-based

- As-consistent-as-possible (ACAP). Gao et al. TVCG. 2019.

- SparseAE. Tan et al. AAAI. 2018.
  - Apply graph convolutional operators with ACAP to analyze mesh deformations.

- VAE Cycle GAN for shape deformation transfer. Gao et al. TOG. 2018.