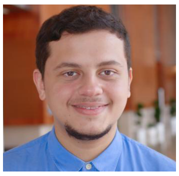
Voint Cloud: Multi-View Point Cloud Representation for 3D Understanding

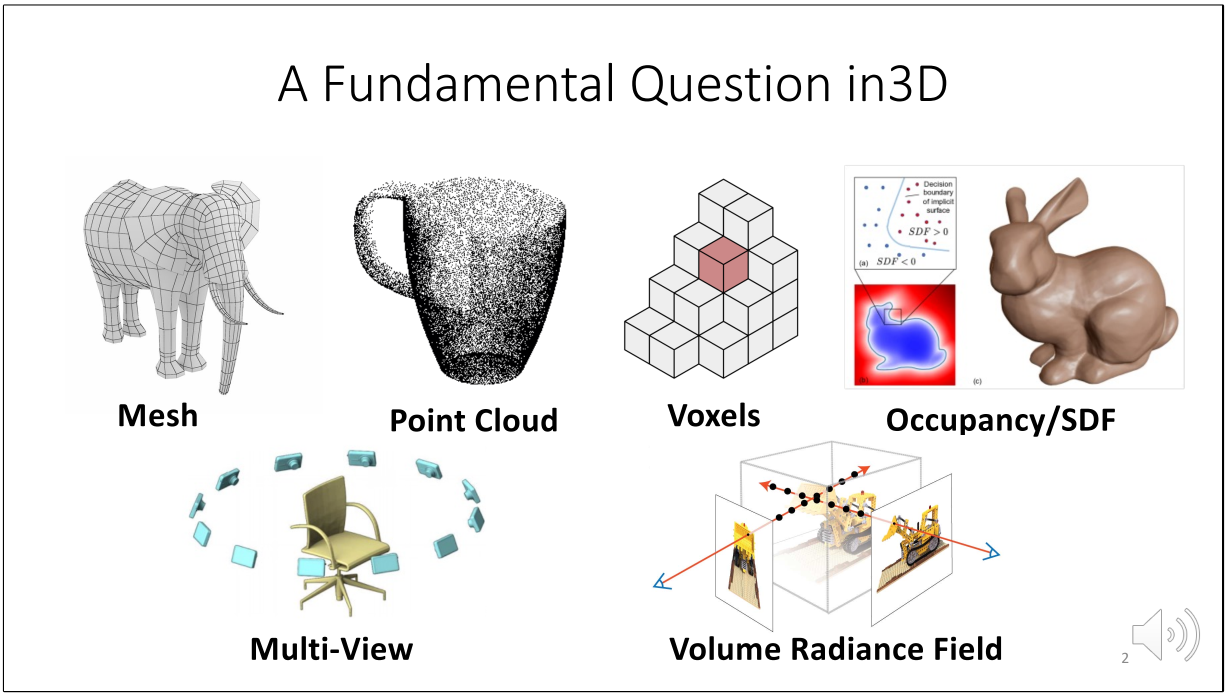
ICLR 2023

Abdullah Hamdi , Silvio Giancola, Bernard Ghanem

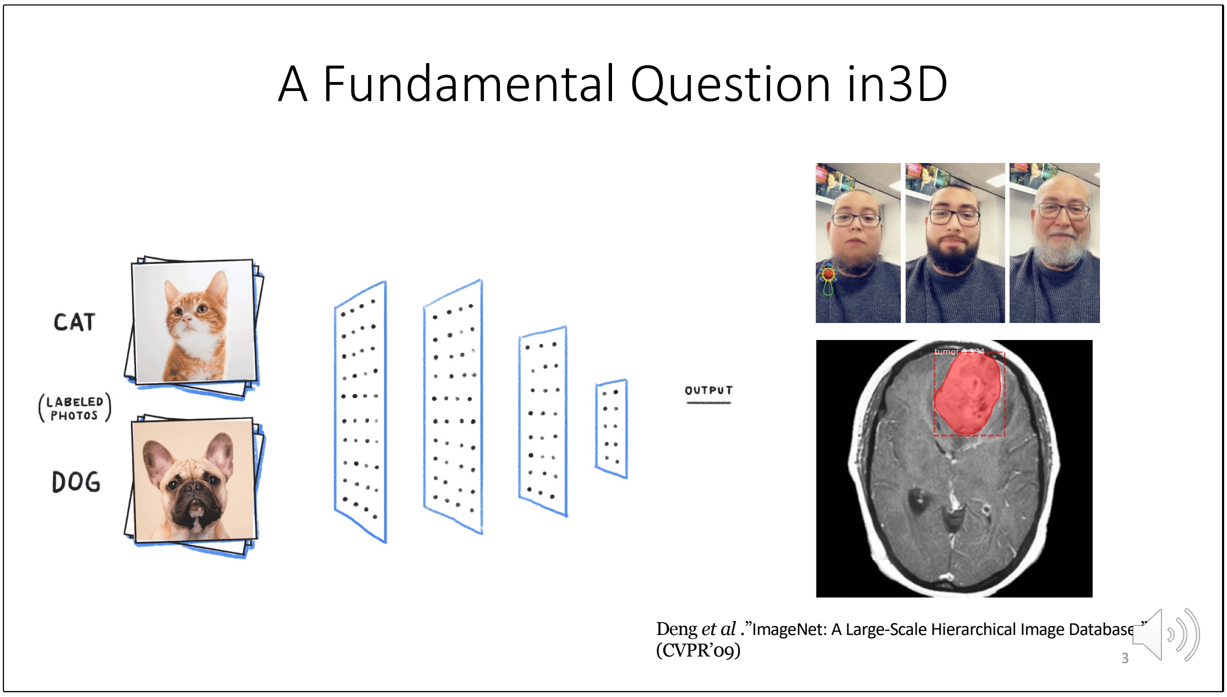
  

KAUST ICLR IVUL

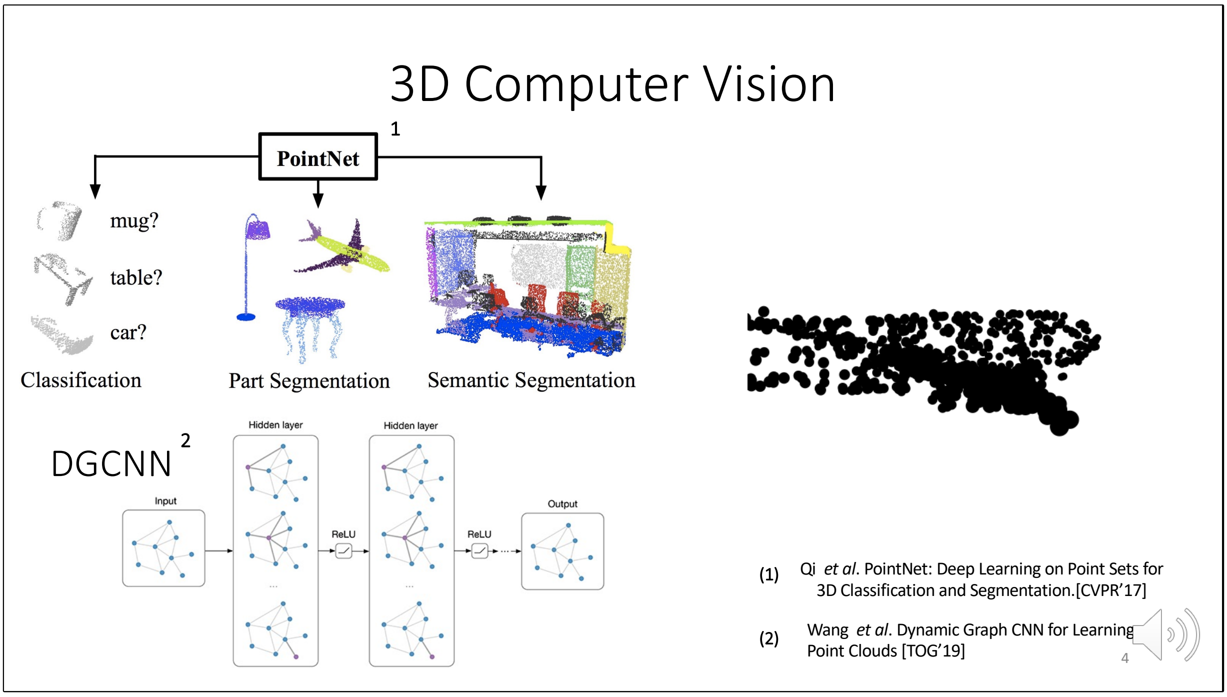
I am Abdullah Hamdi, and I am presenting our ICLR paper : Voint cloud ….



A unfdamental question in 3D vision and graphics is how ro represent 3D data, these includes point clouds , meshes , voxels , implicit coordinate function , multi view and , and lastly volume implicit (nerfs). Ususally this depends on the pplication

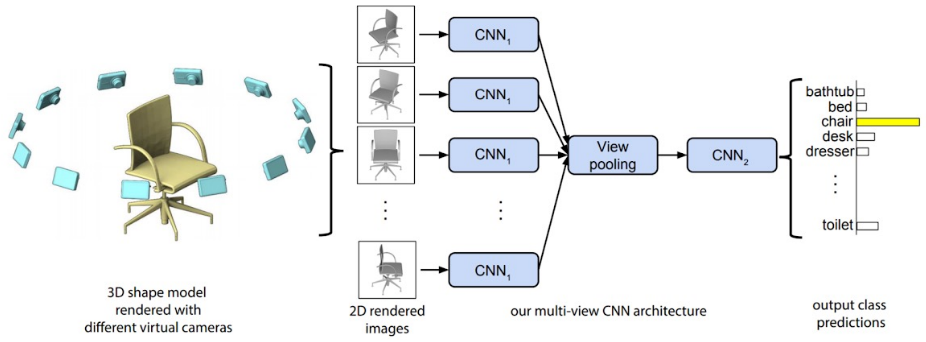


But with the success of 2D deep learning and the wide adoption of deep learning in 3D vision and graphics , emphasized the importance the data structure used to represent 3D



For 3D computer vision , 3D neural netwroks can operate directly on 3D data that are widelky available like 3d point clouds, and has shown success along this direction for many 3D applications like classification and segmentation

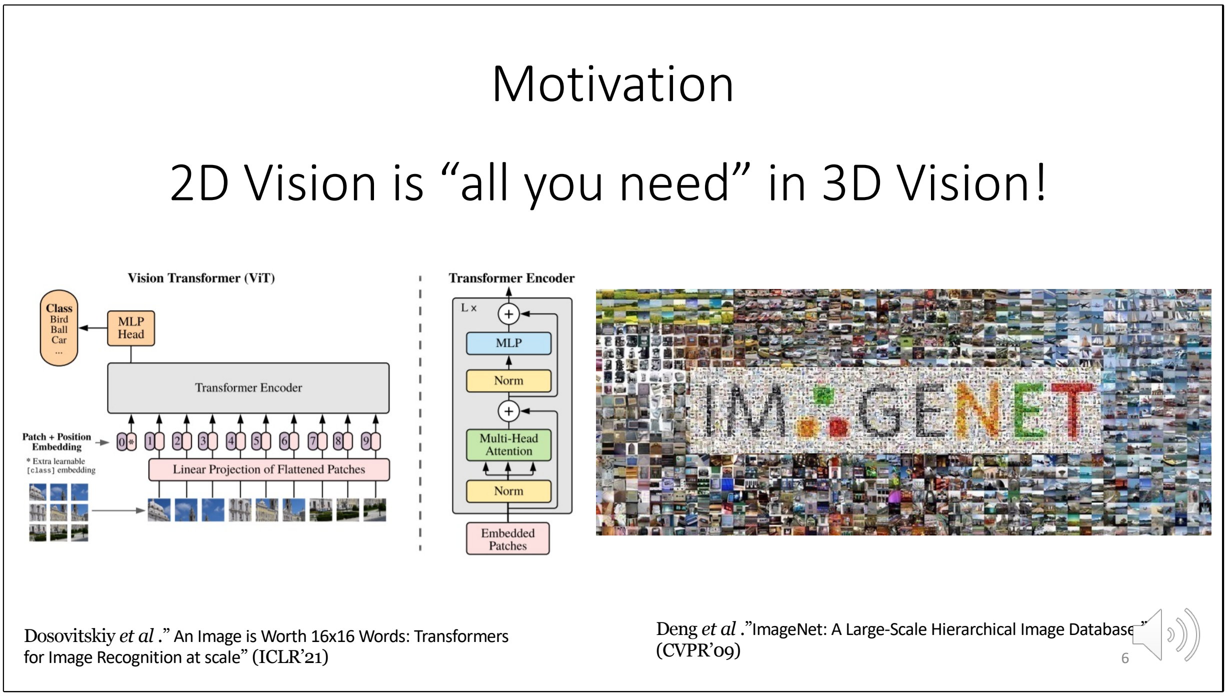
3D Computer Vision



Su, *et.al* ."Multi-view Convolutional Neural Networks for 3D Shape Recognition" (ICCV'15)

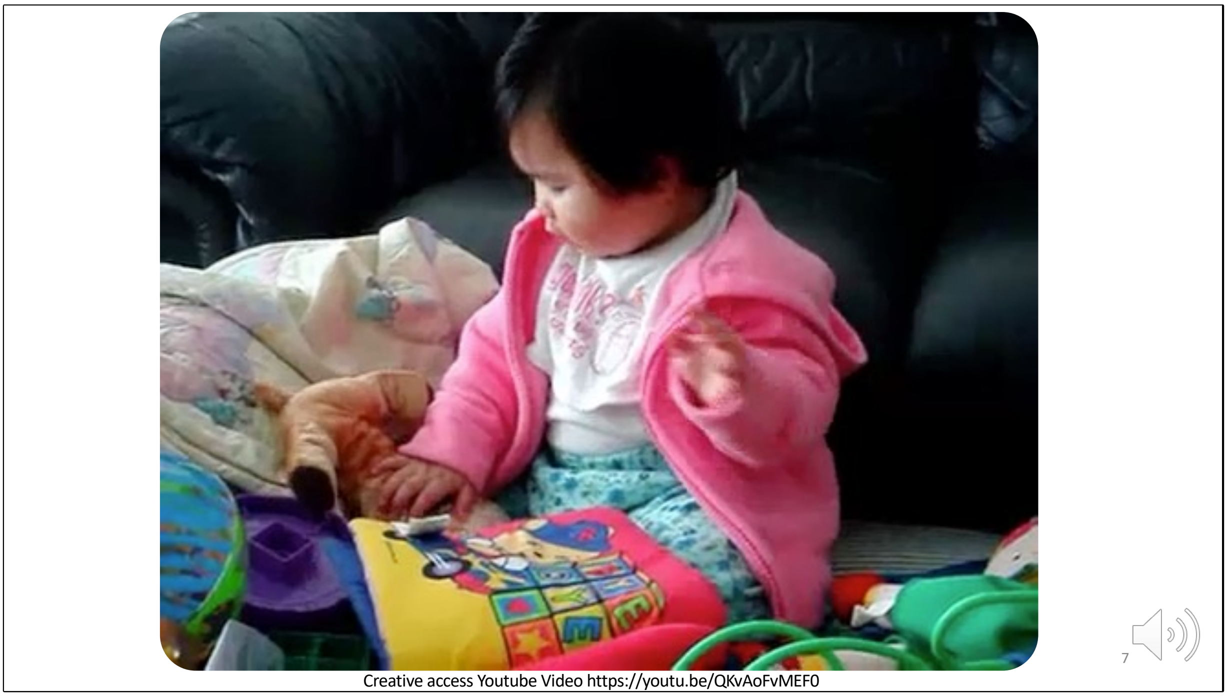
5

The other way is the indirect approach for 3D vision by projecting the 3D data into images in a multi-view approach and then processing the images in a standard 2D pipeline.



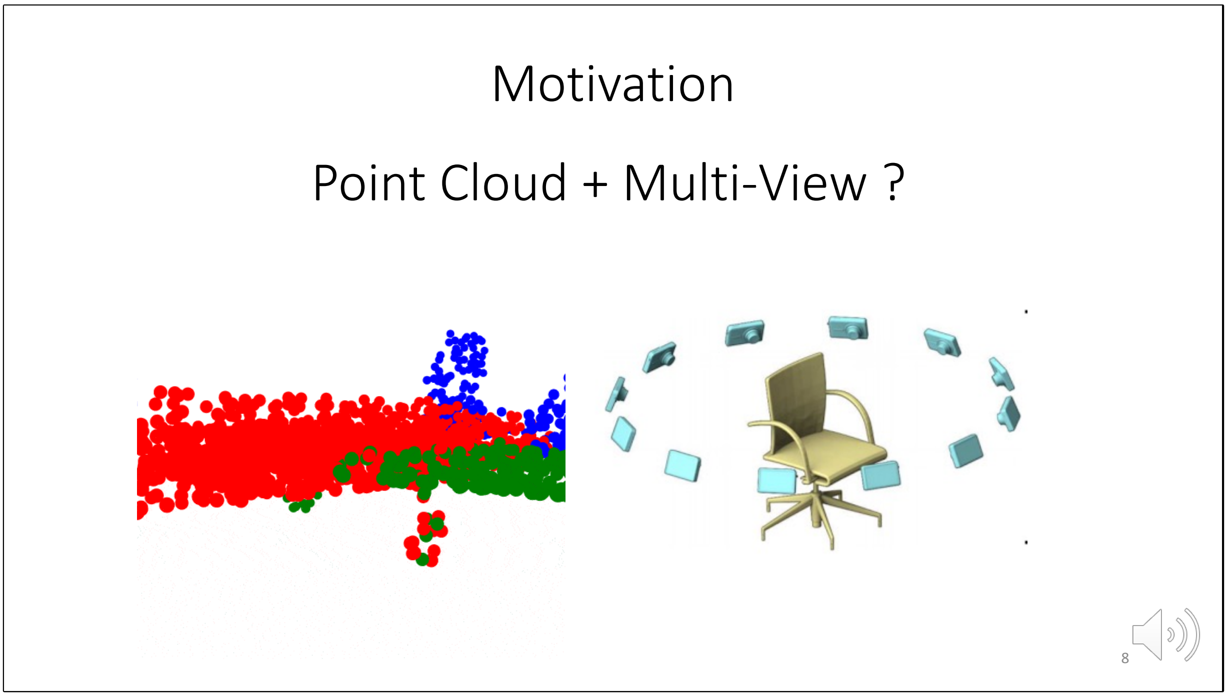
The benefits of Multi-view at the time were clear

✓ **Leveraging the 2D computer vision architectures and methods (eg, CNNs)**
✓ **Leveraging large labeled and diverse 2D image datasets (eg, ImageNet)**



Creative access Youtube Video https://youtu.be/QKvAoFvMEF0

This indirect approach is similar to humans. We dont have 3D sensors. We are naturally looking into objects from differnet angles. We rely on the images projected to our eyes to identify the 3D world

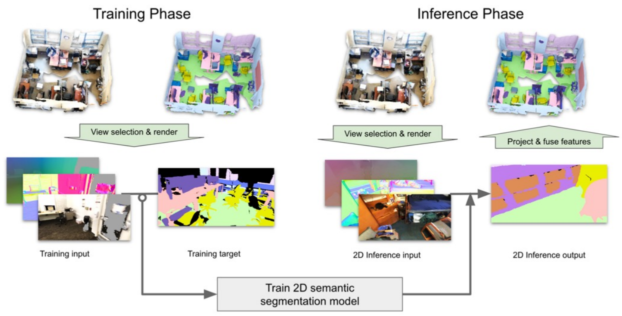


One issue arises when trying to combine widely available 3D point clouds with multi-view(especially for segmentation) with proper per-view aggregation on the point level

Motivation

Mean Fuse (ECCV'20)

Label Fuse (IROS'19)



Kanezaki *et al* ." RotationNet for Joint Object Categorization and Unsupervised Pose Estimation from Multi-View Images" (ECCV'20)

Brian *et al* ." LDLS: 3D Object Segmentation thro[illegible] Label Diffusion from 2D Images" (IROS'19)

9

Previous works used heuristics like mean pooling the features at the point level pr diffusing the labels directly

Motivation

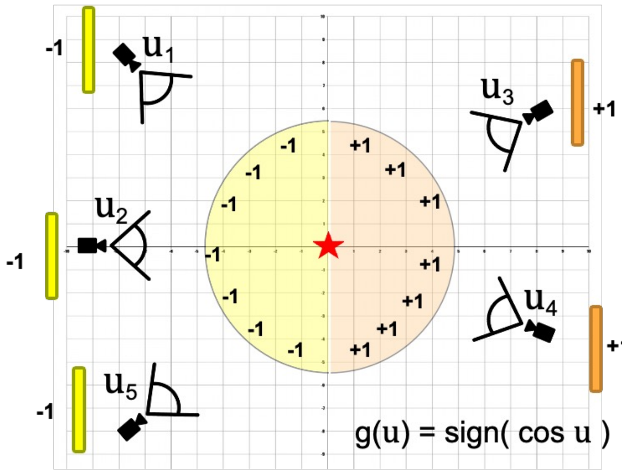
1. **Ignore 3D geometric information**

2. **Depend on the viewing setup**

10

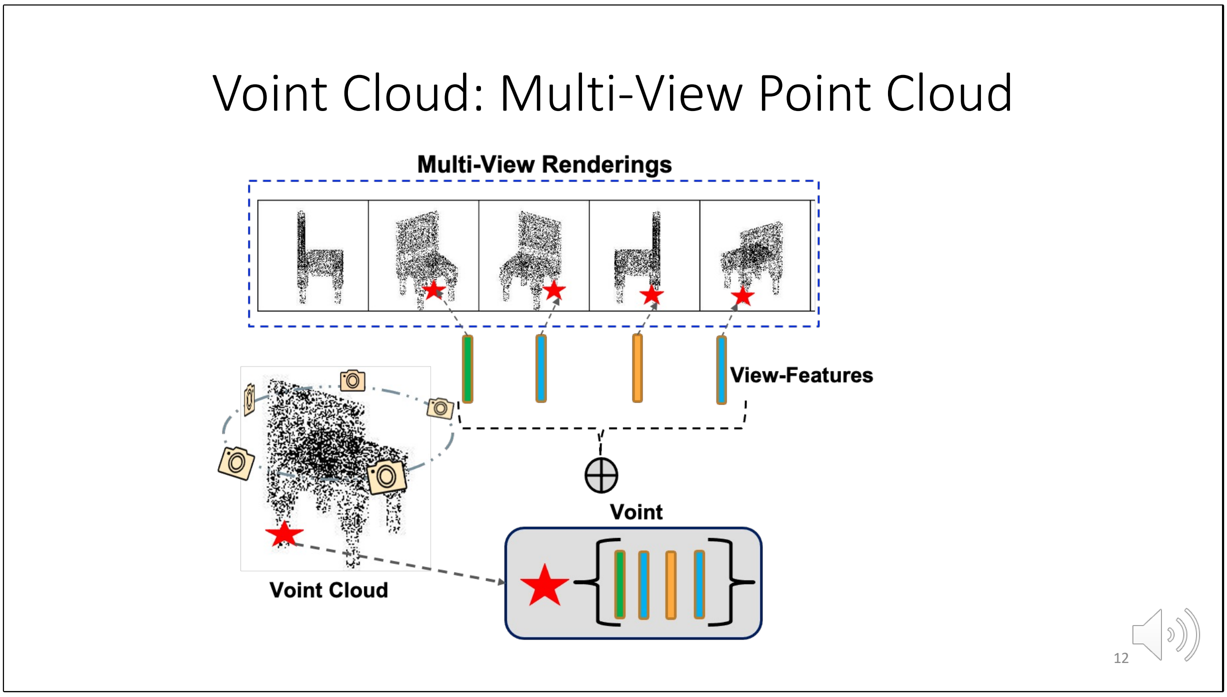
Suuh heuristics of MV + point cloud ignore 3D geometry , Depends on the viewing setup and can lead to fooling views

Motivation

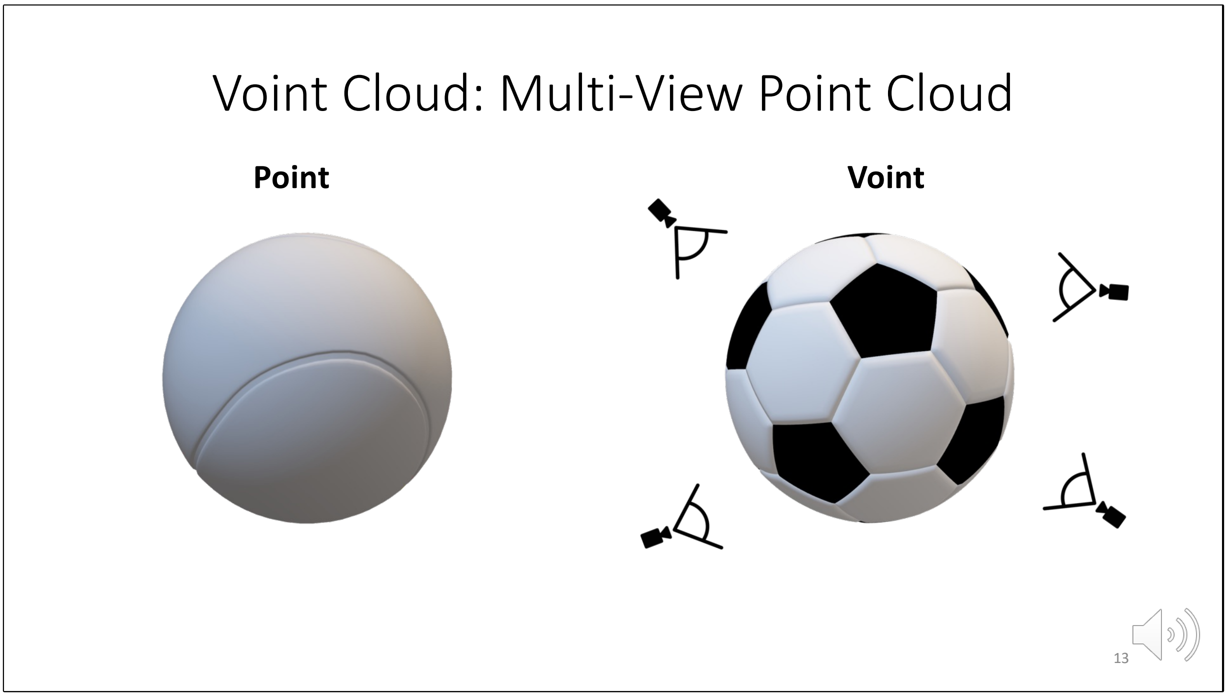


11

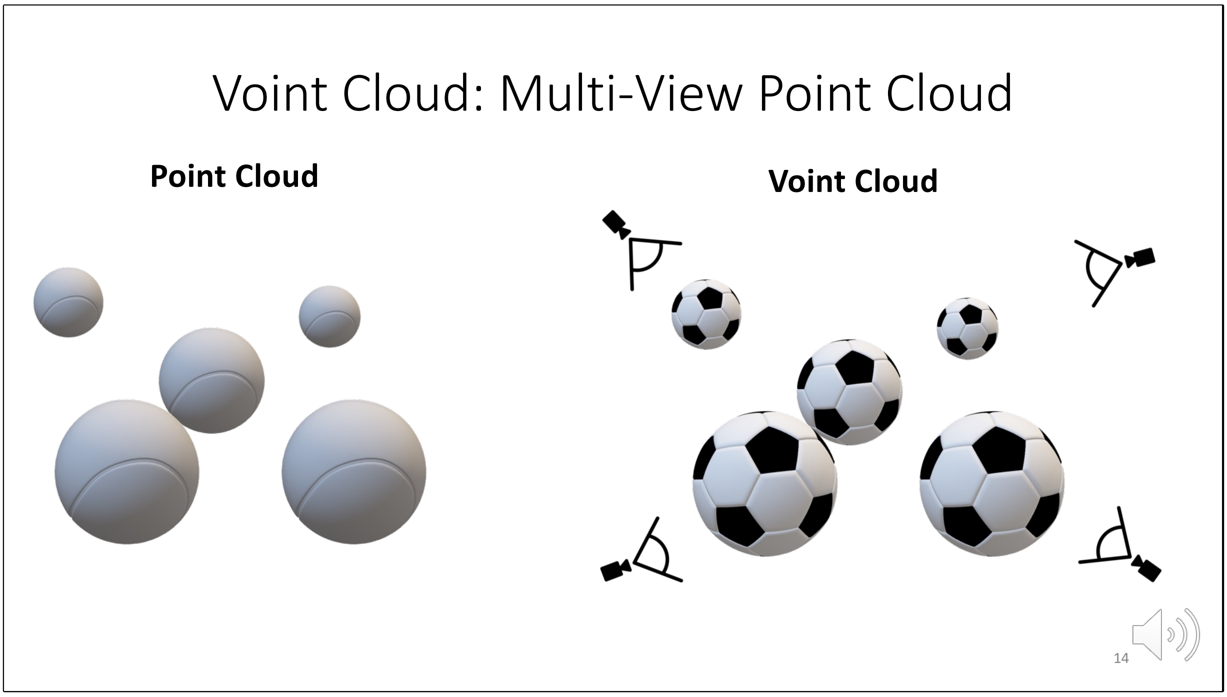
In this toy 2D example, we show that for the same point at the center, different views can give different values and averaging, max the vluaes lose the structure of the underlying function defining



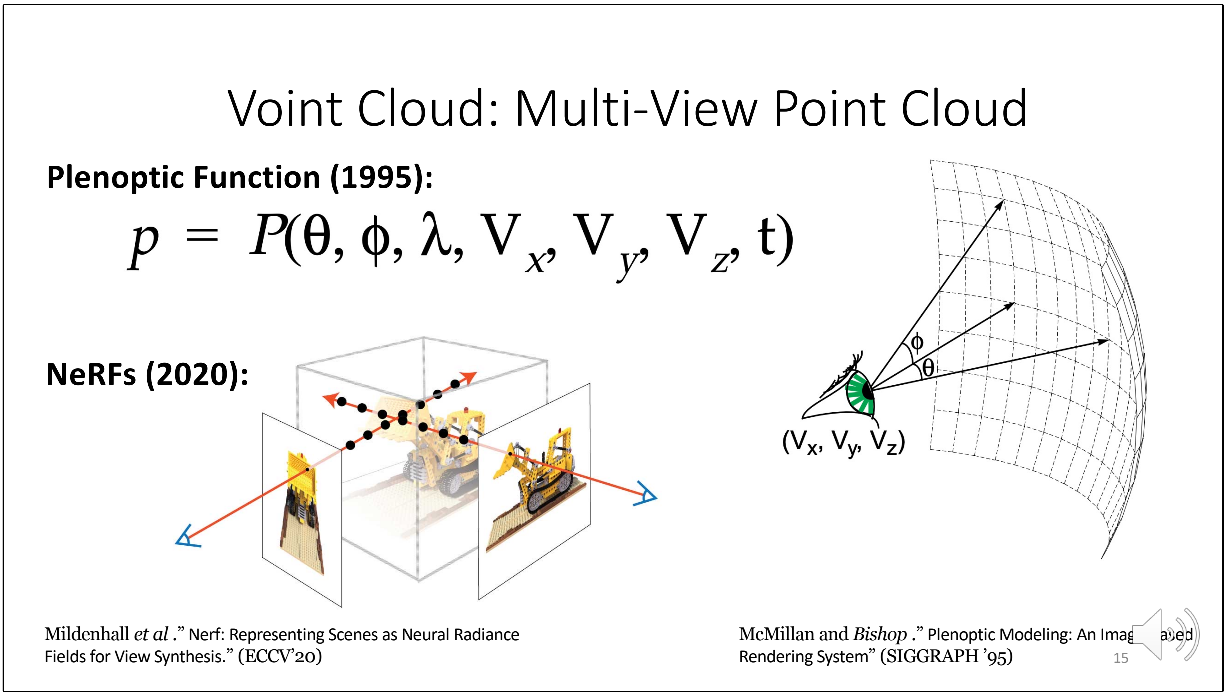
We propose the multi-view point cloud (Voint cloud), a novel 3D representation that is compact and naturally descriptive of view projections of a 3D point cloud. Each point in the 3D cloud is tagged with a Voint, which accumulates view-features for that point.



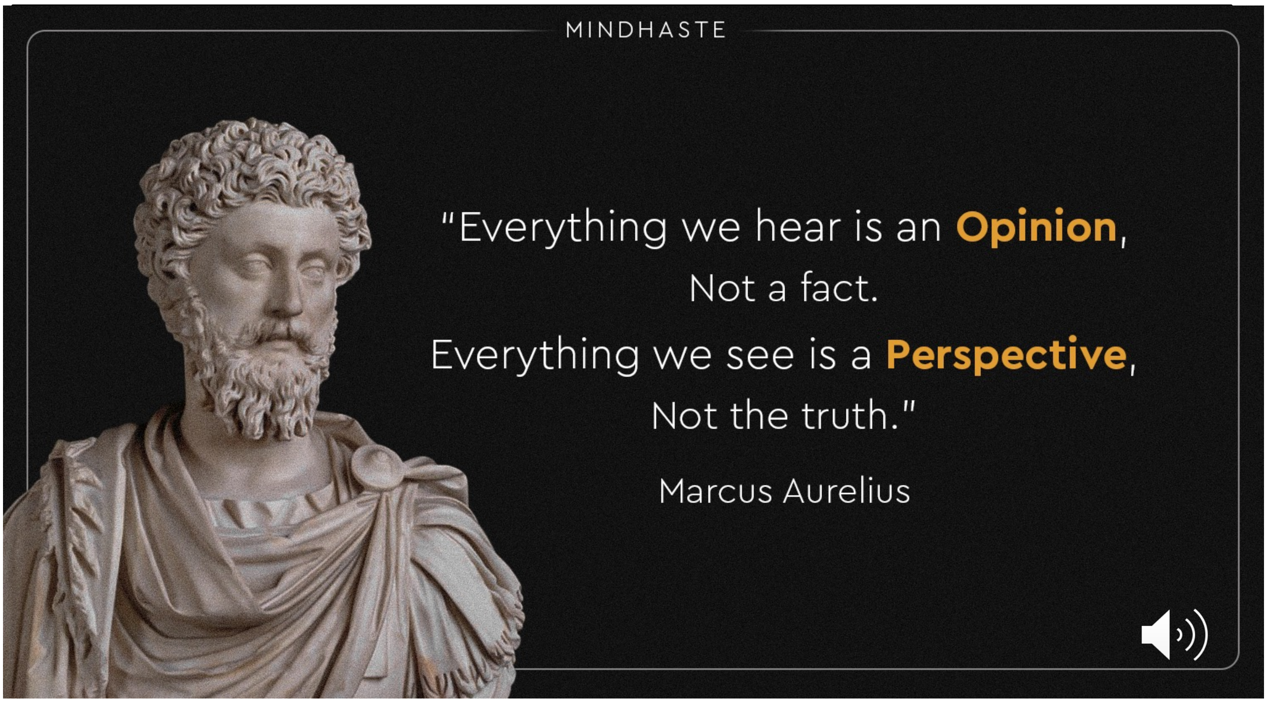
The core assumption in our Voints is that points have a variable value based on the viewing direction , while previous methods assume fixed values for point in point clouds



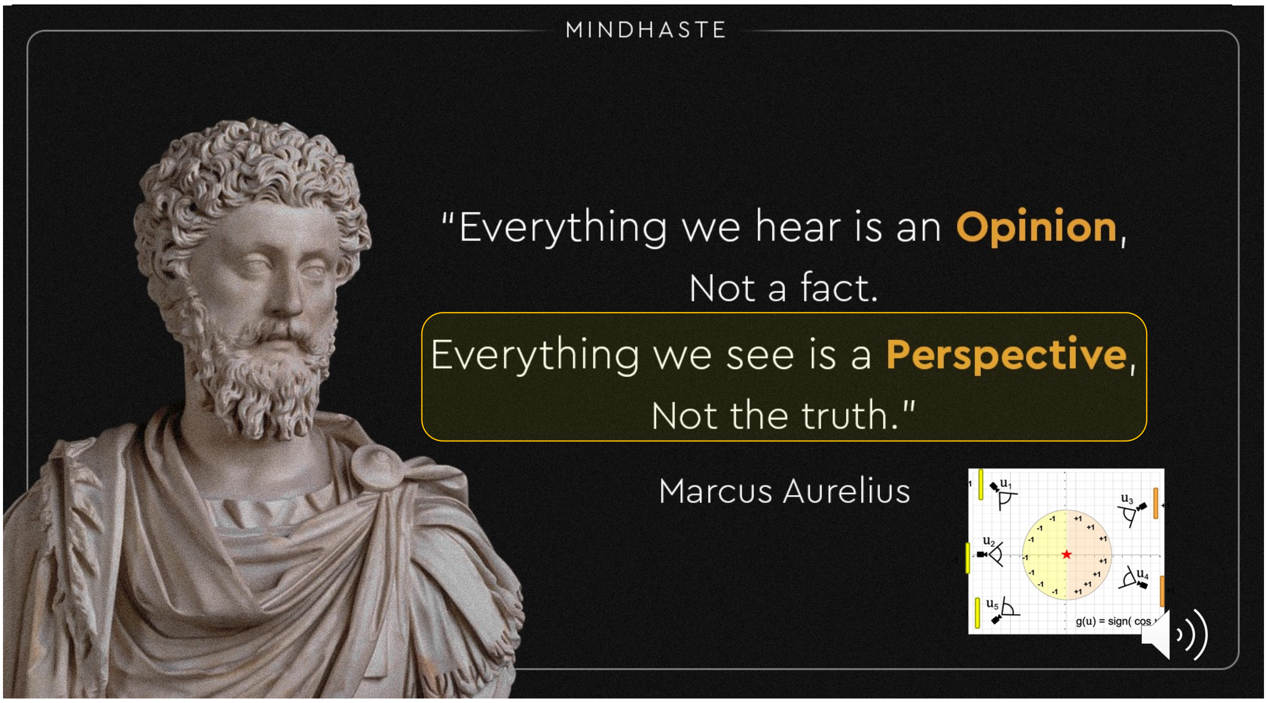
And these view are shared across all the voints



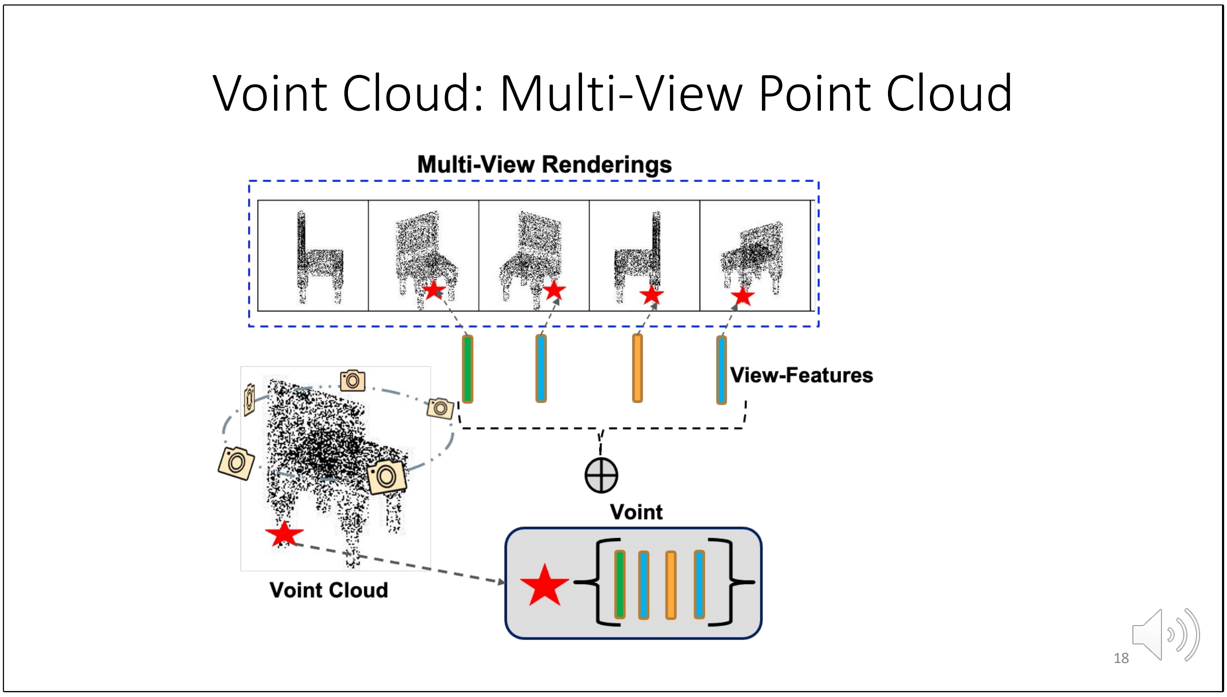
The idea of view dependency is not entirely new. The plenoptic functions in 19995 used them to describe the world from any viewing angle. Nerfs in 2020 usd them to describe radiance fields in neural volume rendering



And even before , Marcus auralius the great roman emperor and philoaspher has a famous quote . Everything we hear is opnion and not fact



Everything we see is a presepctive not the truth … basically , what we see its just one view-point of the underlying truth



So this is basically our Voint cloud representation , a set of voints where each Voint is a set of view-features for the corresponding Point. Note that not all points appear from all the Views and hence each Voint has a different number of view-features

3D Representations Comparison

3D Representation	Explicitness	View-Dependent	Main Use	Memory	3D Descriptiveness
3D Point Cloud	Explicit	✗	3D Understanding	Low	Medium
Multi-View Projections	Implicit	✓	3D Understanding	Medium	Low
Voxels	Explicit	✗	3D Understanding	High	Medium
Mesh	Explicit	✗	3D Modeling	Low	High
Surface Implicit ([39,43])	Implicit	✗	3D Modeling	Medium	High
Volume Implicit (NeRFs [40])	Implicit	✓	Novel View Synthesis	High	Medium
3D Voint Cloud (ours)	Explicit	✓	3D Understanding	Low	Medium

19

In this table from the paper we compare our Voint lcoud to different representations like point clouds , nerfs and voxels

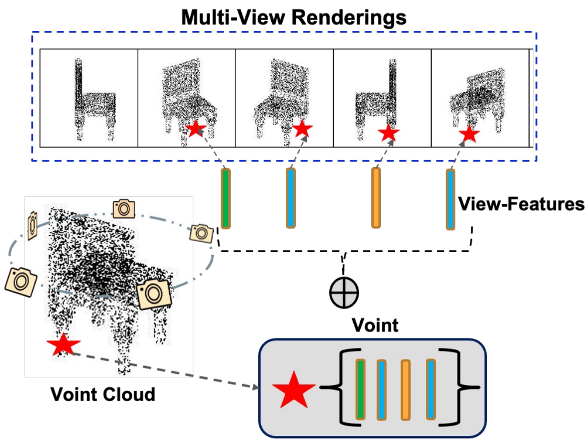
Methodology

20

Voint Clouds

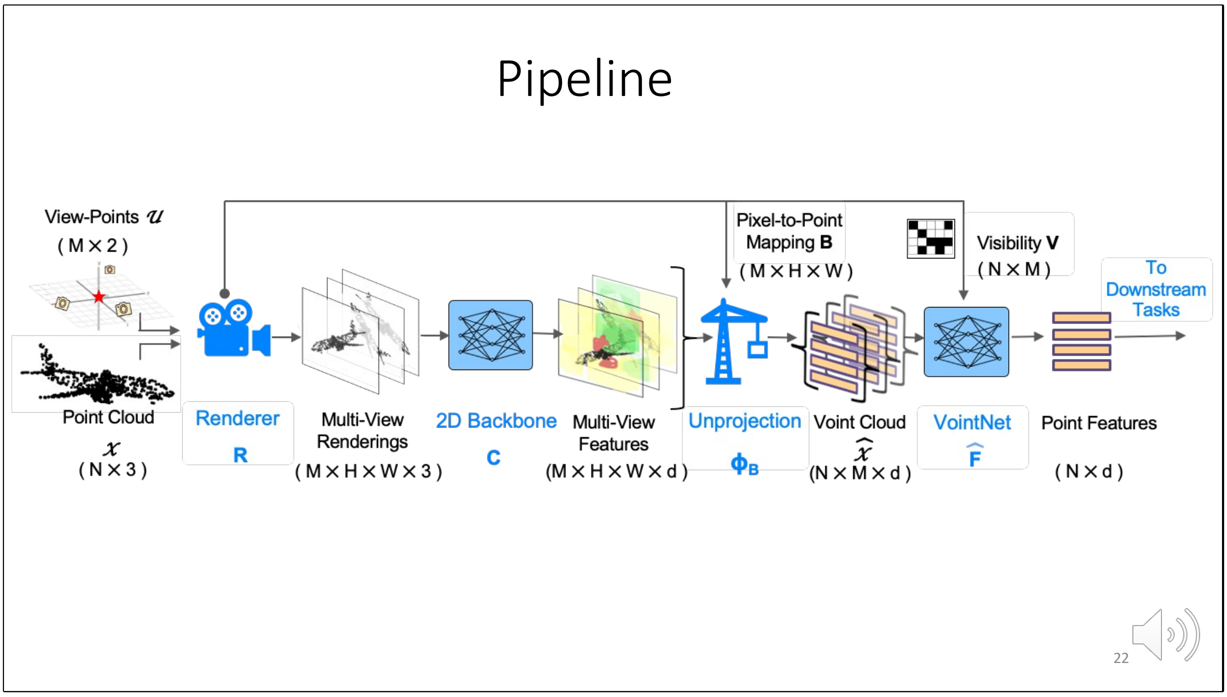
$$\widehat{\mathbf{x}}_i = \left\{\mathbf{g}\left(\mathbf{x}_i, \mathbf{u}_j\right) \in \mathbb{R}^d \mid \mathbf{x}_i \in \mathcal{X}\right\}_{j=1}^{M}$$

$$\widehat{\mathcal{X}} = \left\{\widehat{\mathbf{x}}_i \in \mathbb{R}^{M\times d}\right\}_{i=1}^{N}$$

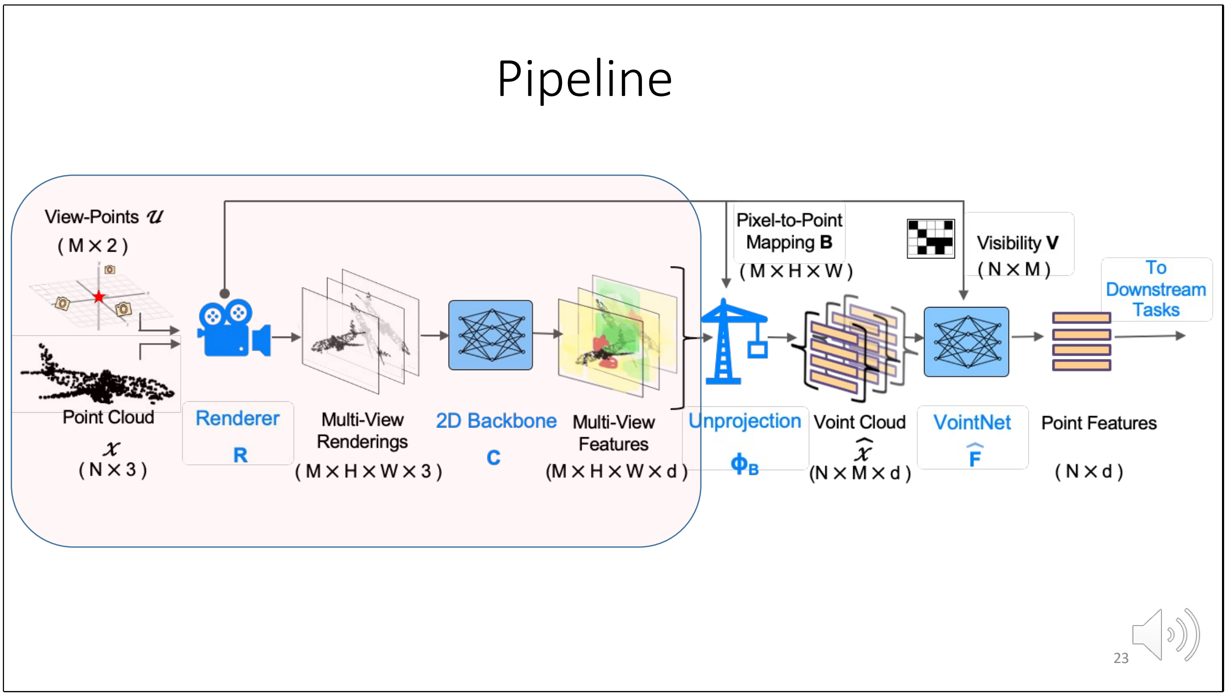


21

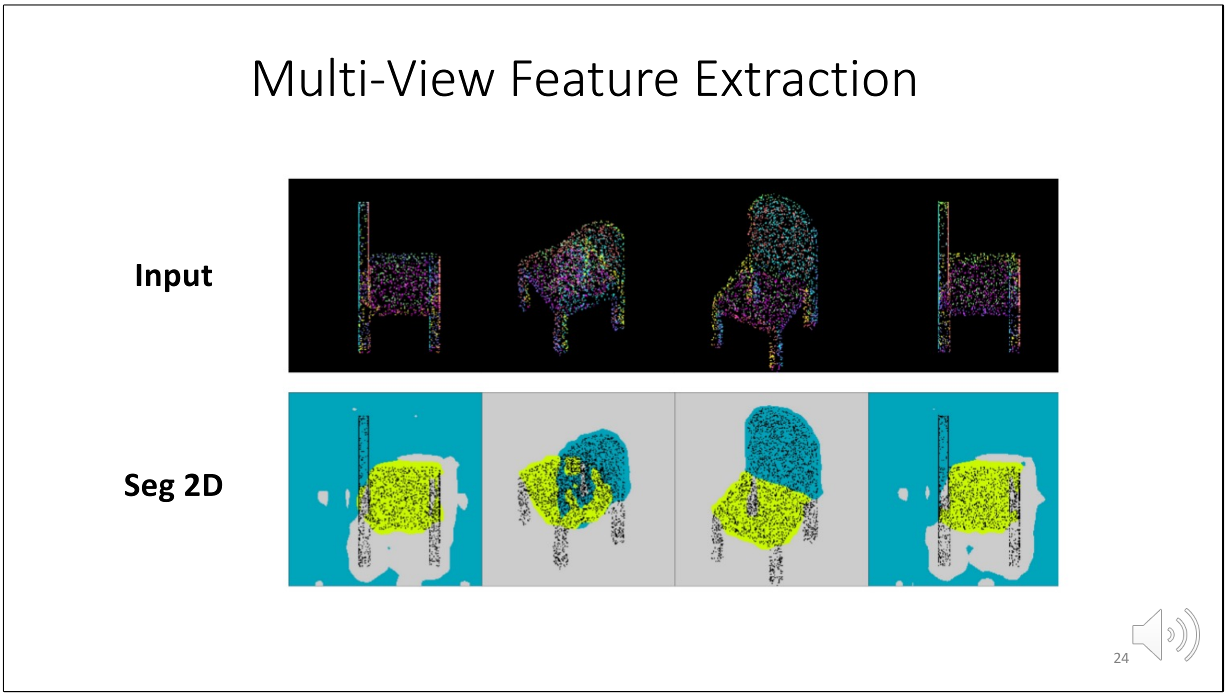
IN our Voint cloud description We said that wach Voint is a set of view-features of the corresponding point … but how do we get these features ?



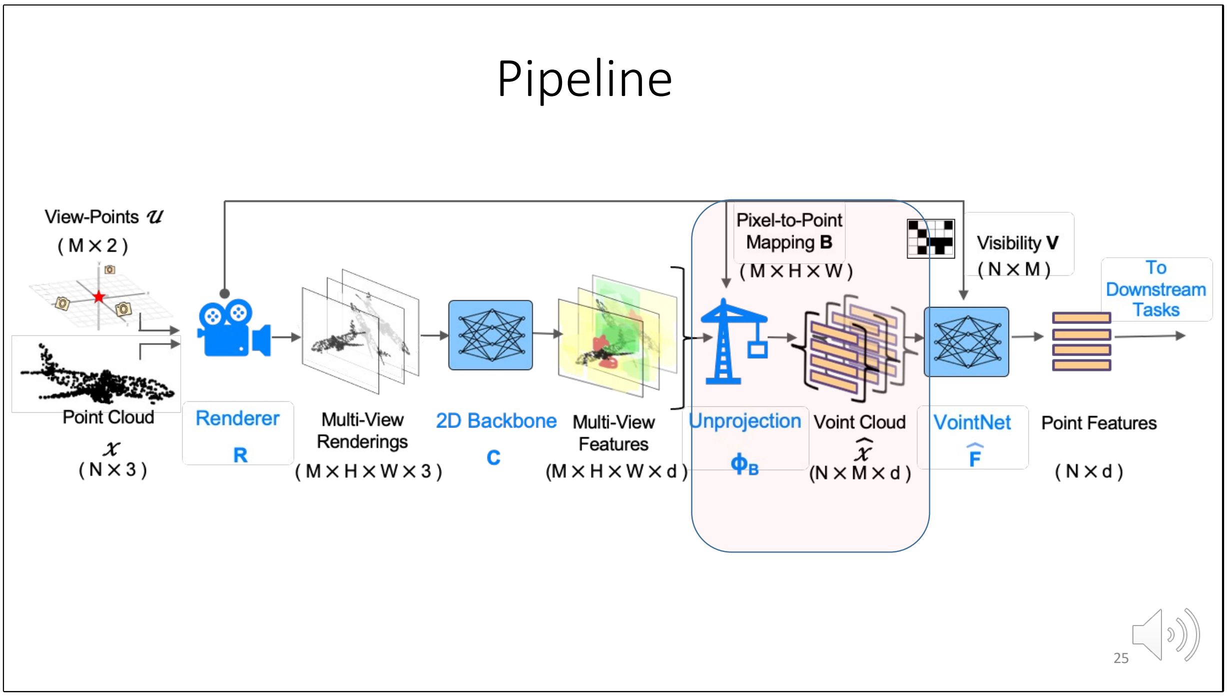
This is our full pipeline



It consists of a arenderer R that render point clouds X from different viewing angles U and the results images are processed by a 2D backbone that extract features per image (these features can be obtained by pre-training the 2D backbone for segmentation or classification)



Here we show examples of outputs of the 2D backbone of the input point cloud renderings



After that the 2d features are unprojected to Voint cloud features using the differentiable module phi_B which uses the mapping B cereate by the renderer that maps every point to pixel

VointNet

VointNet

$$\widehat{\mathbf{F}}(\widehat{\mathcal{X}}) = h_{\mathrm{P}}\left(\mathrm{VointMax}\left(h_{\mathrm{V}}(\widehat{\mathcal{X}})\right)\right)$$

VointMax

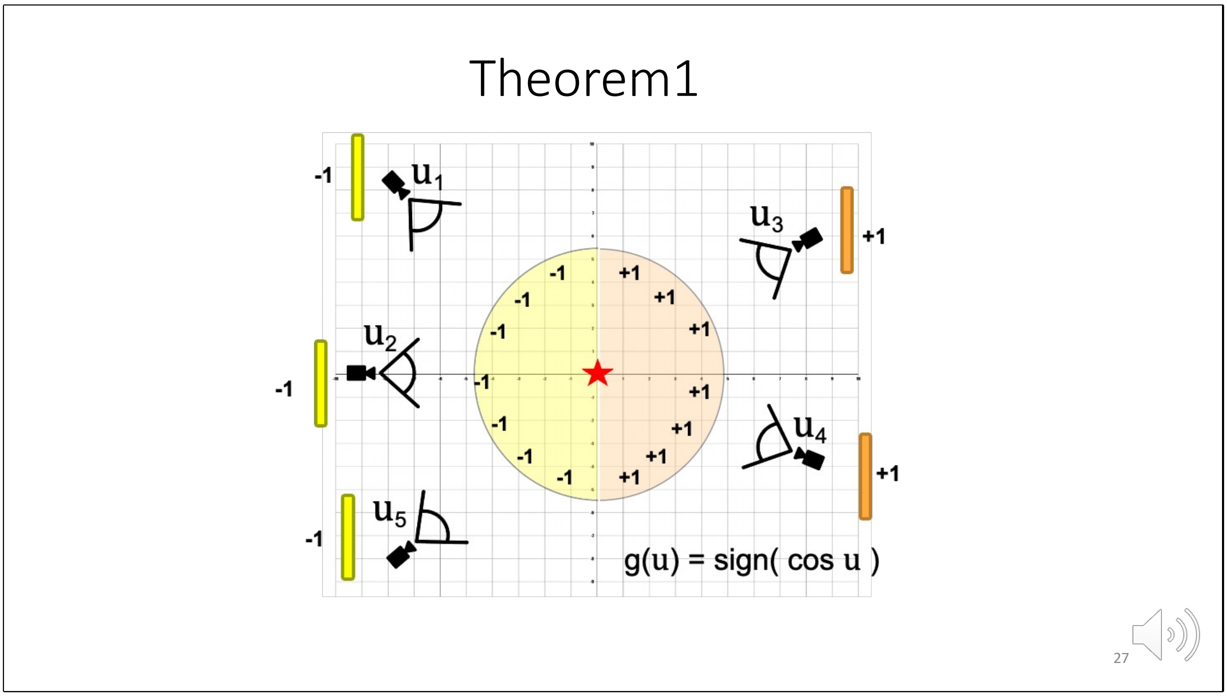
$$\mathrm{VointMax}(\widehat{\mathbf{x}}) = \max_{j} \widehat{\mathbf{x}}_{i,j},\ \forall i, j$$
$$\text{s.t. } i \in 1, 2, ..., N\ ,\ j \in 1, 2, ..., M\ \ , \mathbf{V}_{i,j} = 1$$

VointConv

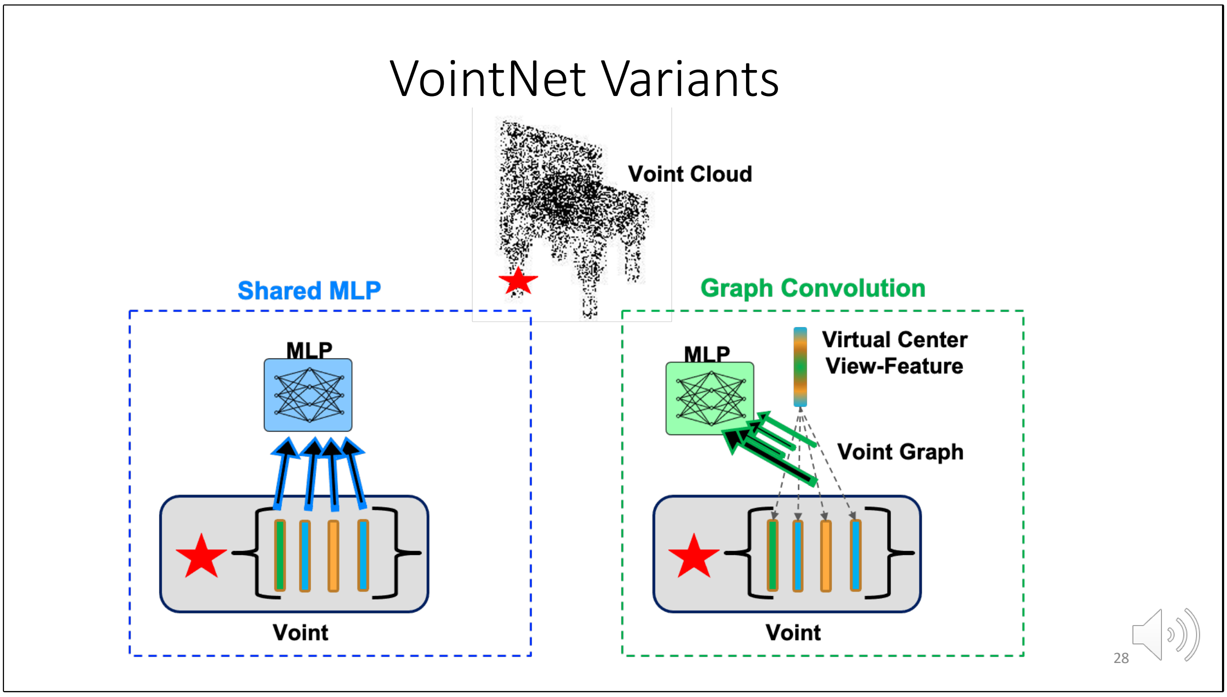
$$\mathbf{h}_{i,j}^{l+1} = \rho\left(\mathbf{h}_{i,j}^{l} \mathcal{W}_{\rho}\right),\ \forall i, j$$
$$\text{s.t. } i \in 1, 2, ..., N\ ,\ j \in 1, 2, ..., M\ \ , \mathbf{V}_{i,j} = 1$$

26

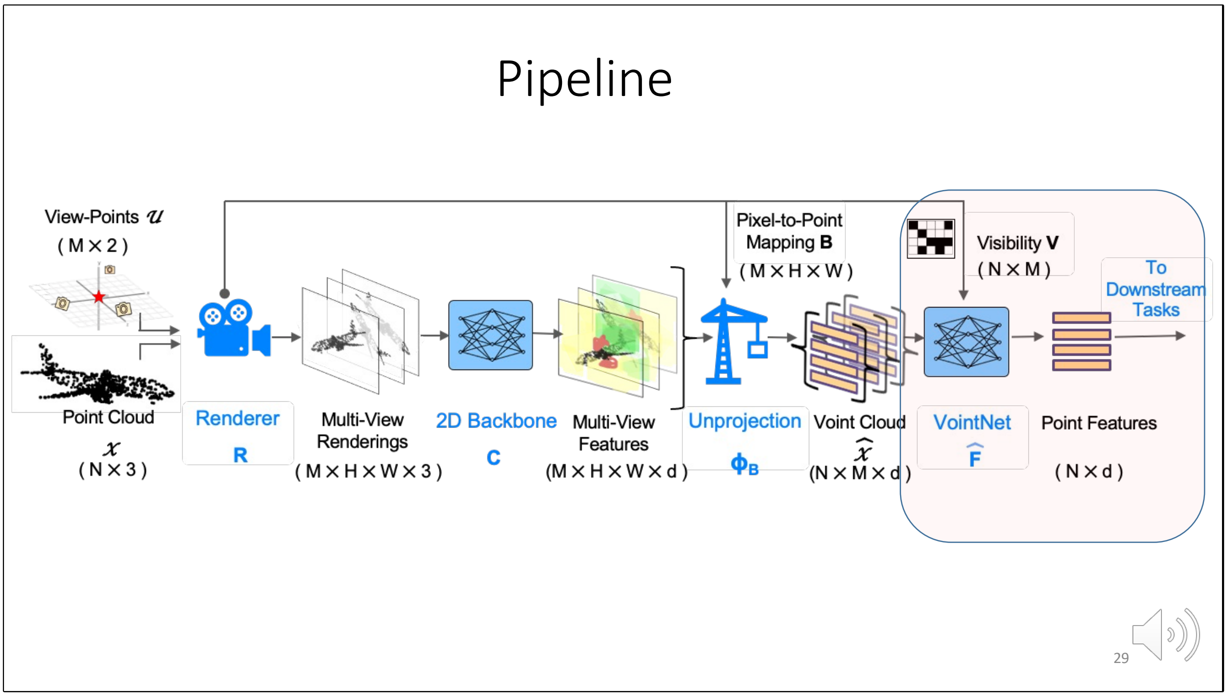
In order to learn on the Voint space we propose Vointnet in the following form . A VointConv followed by a Vointmax where



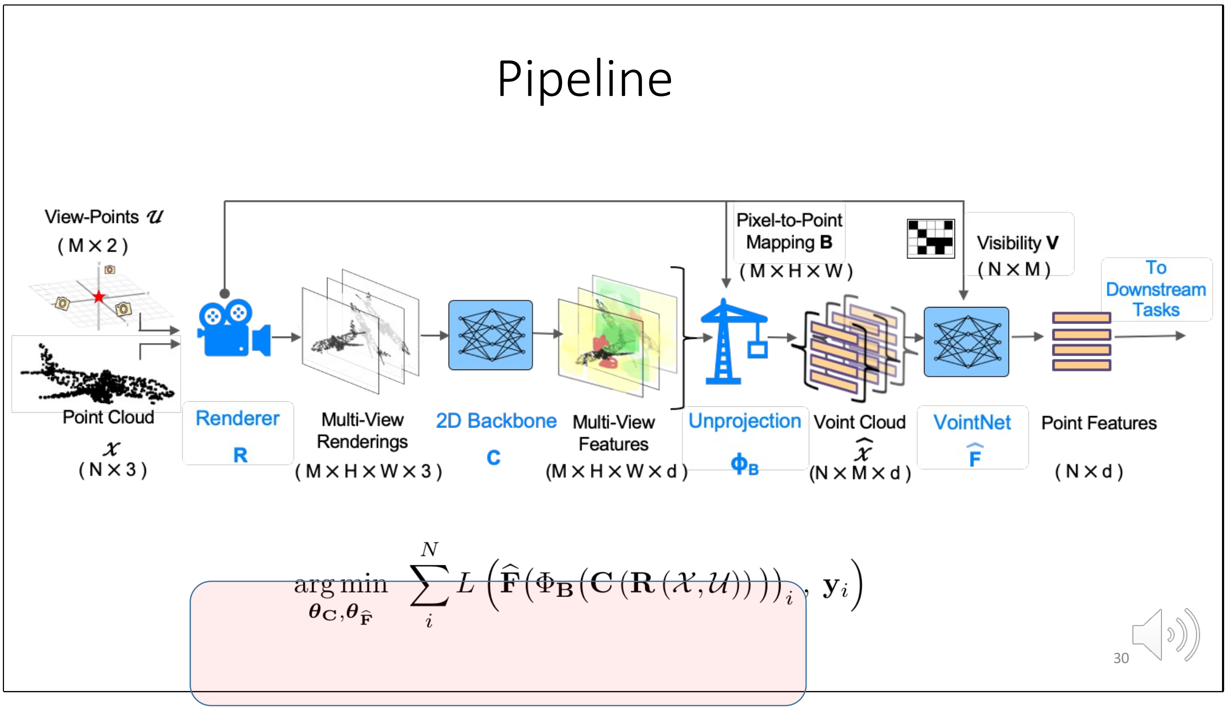
This form of max view features is proven in Theorem 1 in the paper to be a global approximater to any funcrtion on the set of angles U in the 2D case



We can use shared MLP as the Voint conv operation applied independently on all view features and shared wights . Or we can do a graph on the view dimension and define the mlp on that edge features . We define GAT as well



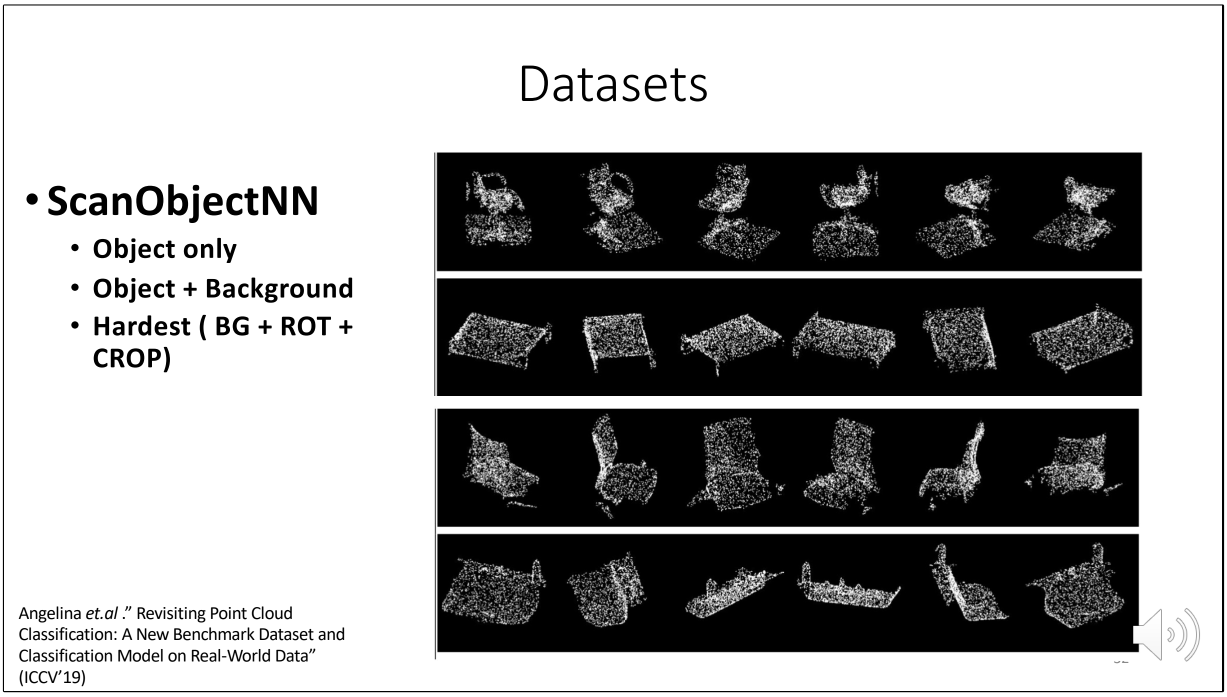
The VointNet outputs point cloud features that are ready for any typical point cloud processing pipeline



The VointNEt pipeline is then trained end-end with focus on the vointnet part since the 2D backbone is pretrained on the task in hands. We learn both F and C for in the loss optimization

Experiments

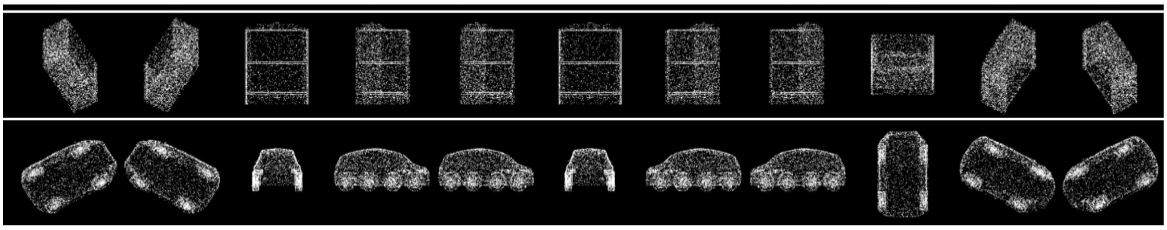
31



Here is a visualization of the dtaasets we used when rendered in our pipeline The first dataset is ScanObjectNN with 2,902 point clouds and 15 classes. It conisist of realistic 3D scans of objects and has 3 Variants

Datasets

- **ShapeNet Core55**



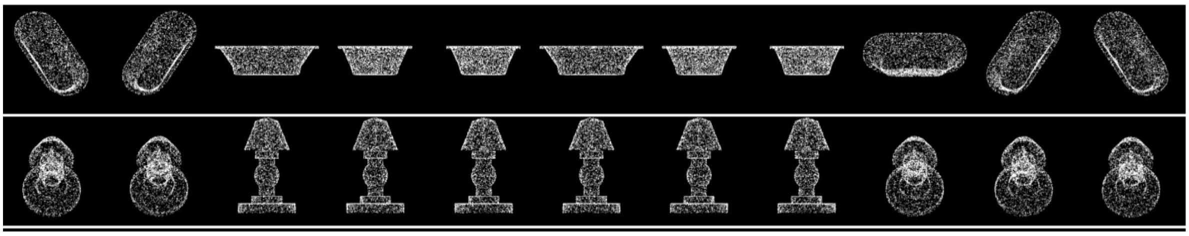
Cheng. *et.al* ." ShapeNet: An Information-Rich 3D Model Repository" (arxiv'15)

33

ShapeNet Core 55 for retrieval

Datasets

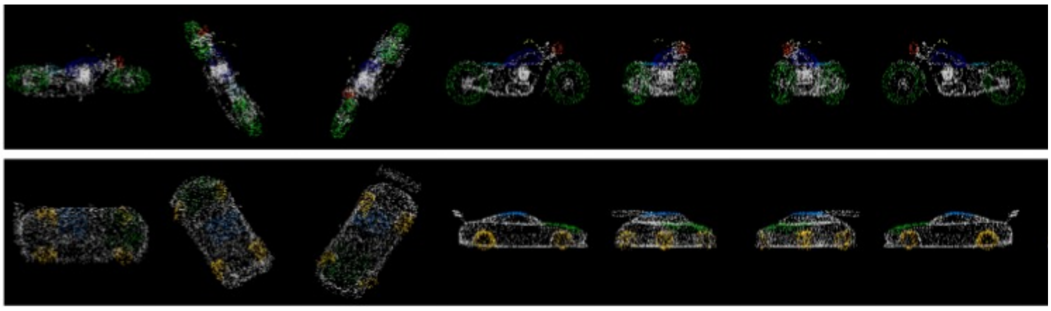
- **ShapeNet Core55**



Cheng. *et.al* ." ShapeNet: An Information-Rich 3D Model Repository" (arxiv'15)

Datasets

- **ShapeNet Parts**



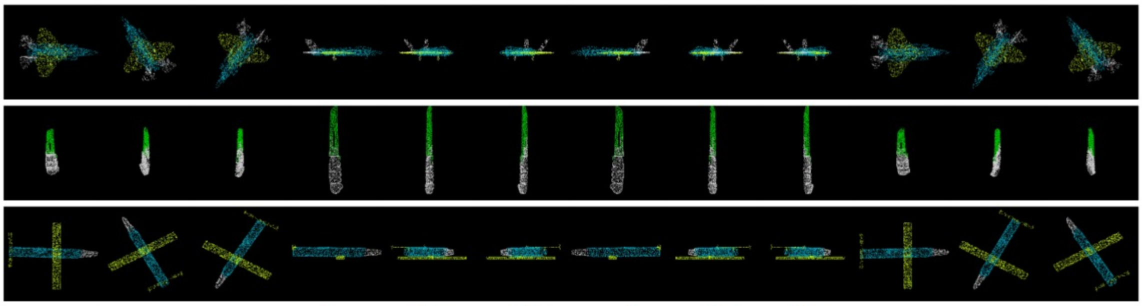
Yi, *et.al* ." scalable active framework for region annotation in 3d shape collections." (TOG'16)

35

Shape Net parts for segmentation . We show the labels with different colors , and thos renderings are used to train 2D segmenter

Datasets

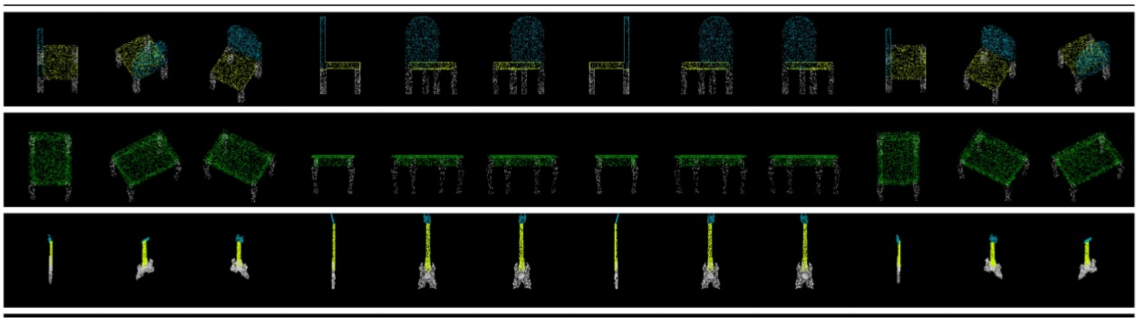
- **ShapeNet Parts**



Yi, *et.al* ." scalable active framework for region annotation in 3d shape collections." (TOG'16)

Datasets

- **ShapeNet Parts**

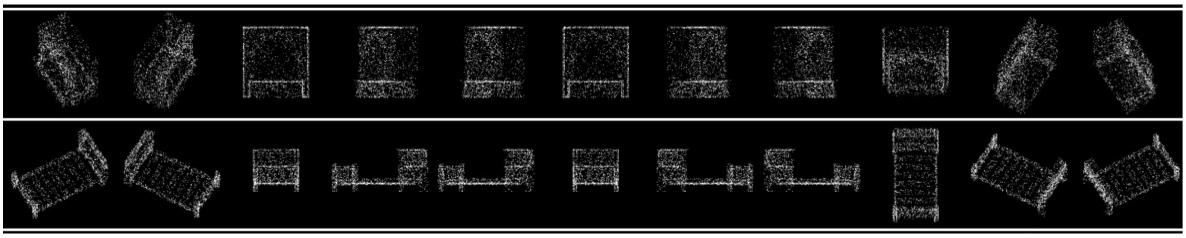


Yi, *et.al* ." scalable active framework for region annotation in 3d shape collections." (TOG'16)

More examples

Datasets

- **ModelNet40**



Wu. *et.al* ." 3D ShapeNets: A Deep Representation for Volumetric Shapes" (CVPR'15)

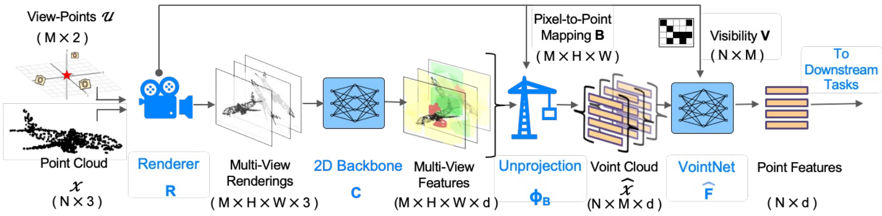
38

. ModelNEt 40 for classification

Details

ViT-B : backbone for classification

DeepLabV3 : backbone for segmentation



Dosovitskiy *et al* ." An Image is Worth 16x16 Words: Transformers for Image Recognition at scale" (ICLR'21)

Chen *et al* ." Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation" (ECCV18)

39

The backbone we used for classification is Vit-B and for segmentation we used DeeplabV3

Lets have a look at the results

3D Point Cloud Classification

ScanObjectNN

Method	Data Type	Classification Overall Accuracy OBJ_BG	OBJ_ONLY	Hardest
PointNet (Qi et al., 2017a)	Points	73.3	79.2	68.0
SpiderCNN (Xu et al., 2018)	Points	77.1	79.5	73.7
PointNet ++ (Qi et al., 2017b)	Points	82.3	84.3	77.9
PointCNN (Li et al., 2018)	Points	86.1	85.5	78.5
DGCNN (Wang et al., 2019c)	Points	82.8	86.2	78.1
SimpleView (Goyal et al., 2021)	M-View	-	-	79.5
MVTN (Hamdi et al., 2021)	M-View	92.6	92.3	82.8
VointNet (ours)	Voints	**93.7**	**94.0**	**85.4**

41

On the reslistic ScaNobjectNN dataset we achive SOTA on all three variants

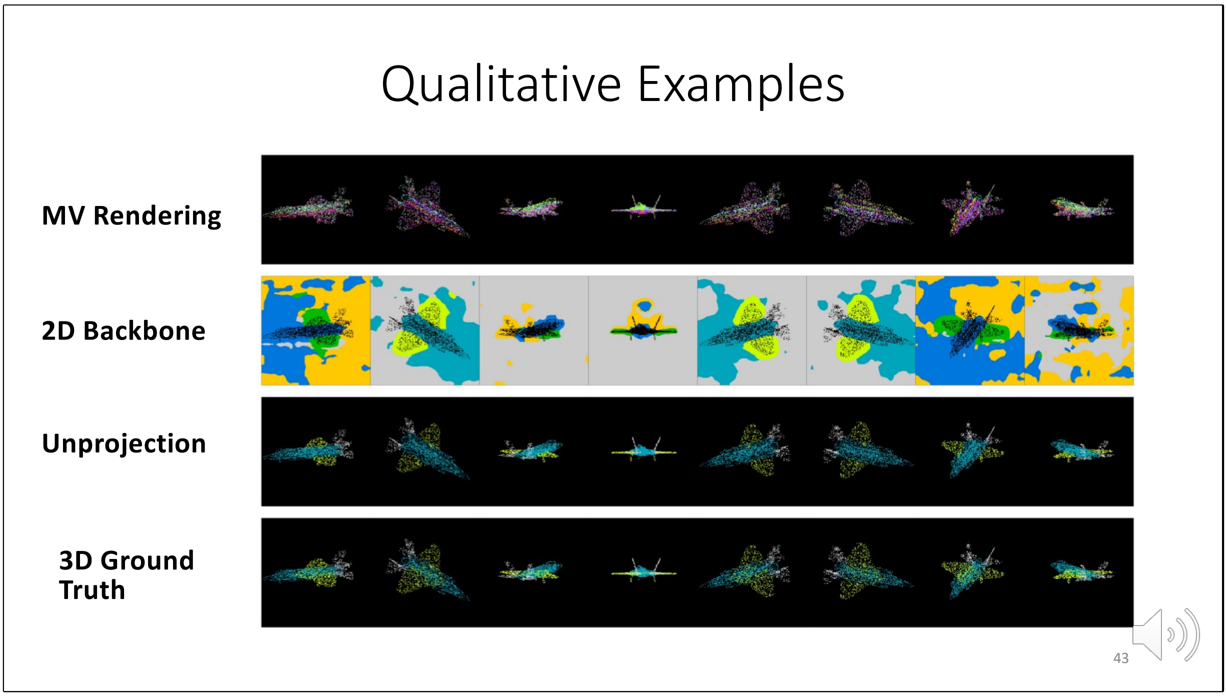
3D Shape Retrieval

ShapeNet Core55

Results	MVCNN (Su et al., 2015)	RotNet (Kanezaki et al., 2018)	ViewGCN (Wei et al., 2020)	MVTN (Hamdi et al., 2021)	VointNet (ours)
ShapeNet Retr. mAP	73.5	77.2	78.4	82.9	**83.3**

42

We achive SOTA on ShapeNet Core 55 retrieval benchmark compared to strong and recent multi-view methods specialized for retrieval



Here we show how the renderings colored with normals and then 2D segmented can be unprojected to 3D predictions and compare them to 3D GT labels

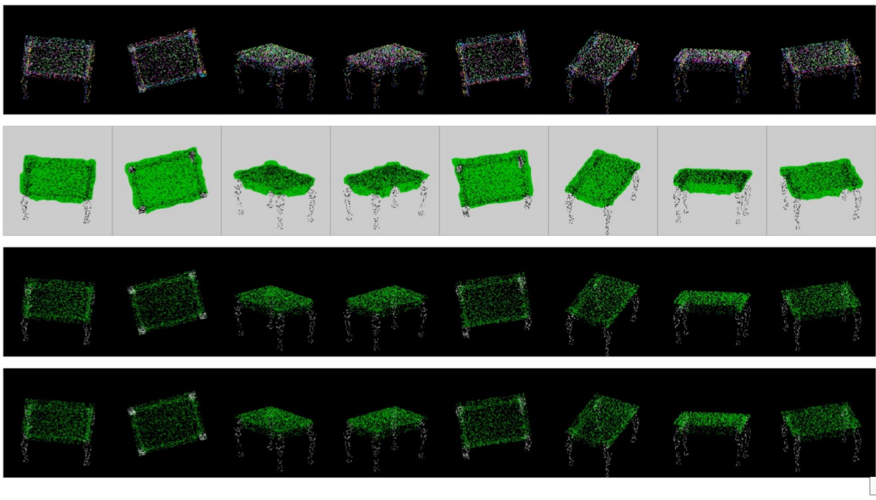
Qualitative Examples

MV Rendering

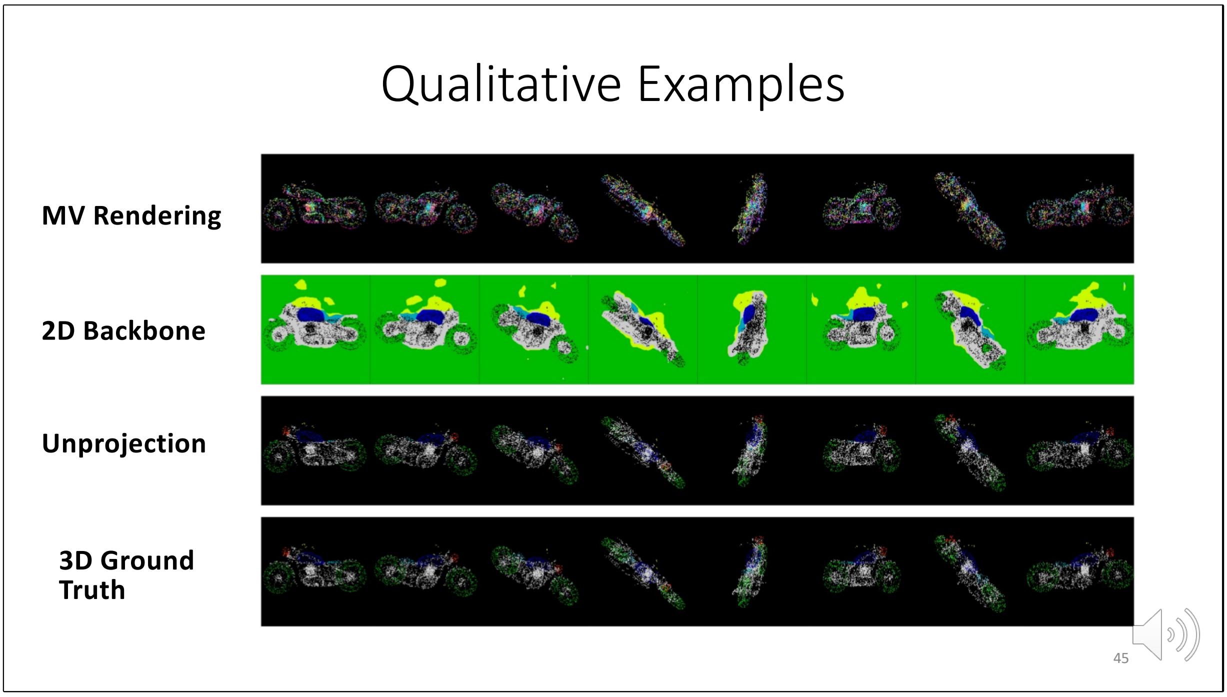
2D Backbone

Unprojection

3D Ground Truth

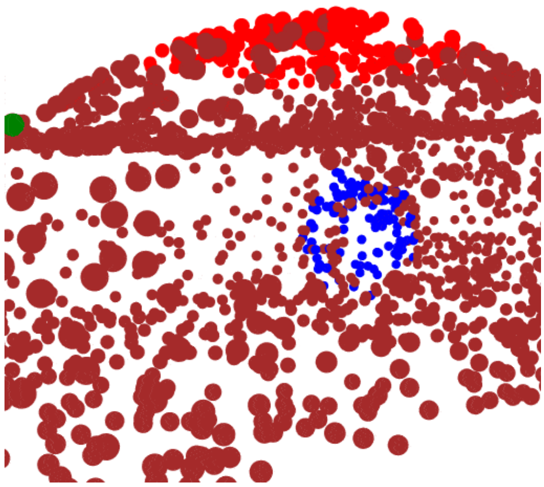


44



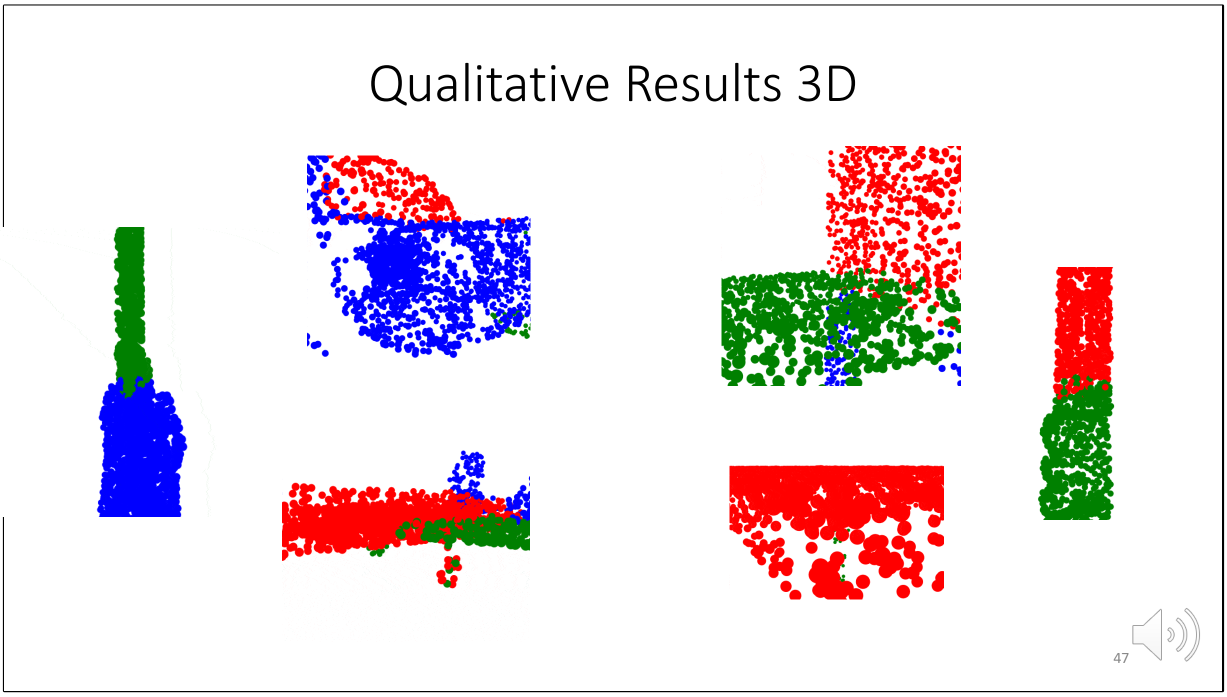
Here we show some qualitative examples of shape retrieval

Qualitative Results 3D

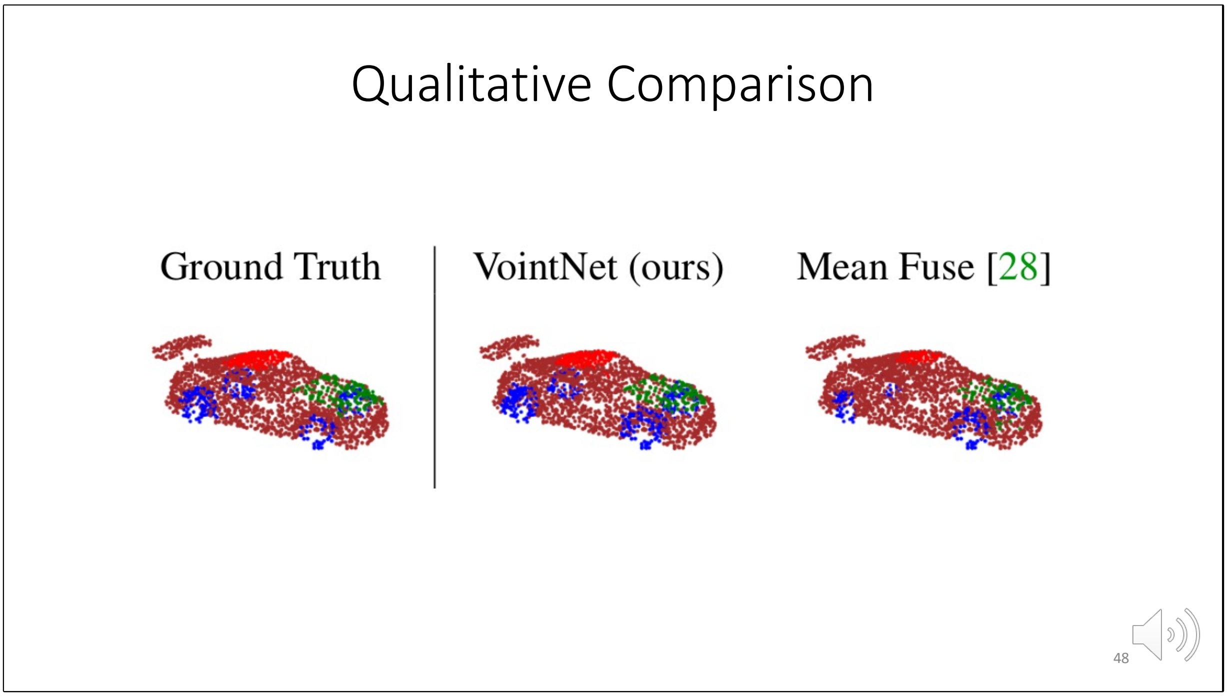


46

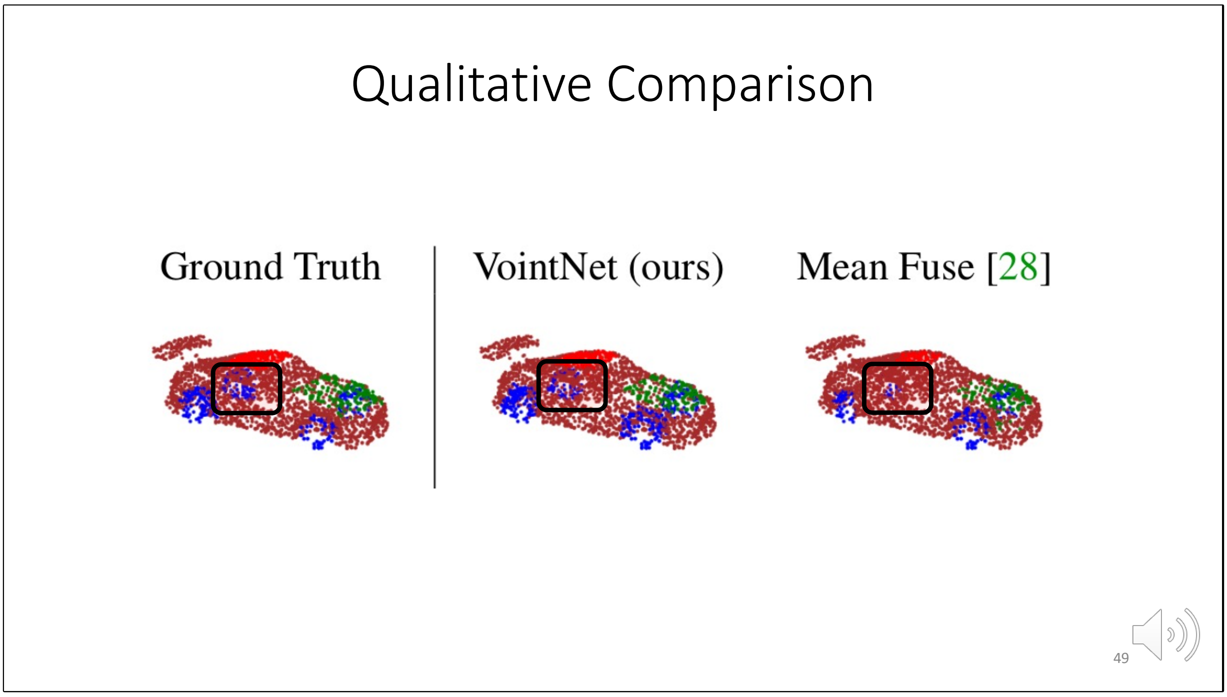
Here we show example of the 3D segmentation from our VointNEt



Here we show example of the 3D segmentation from our VointNEt

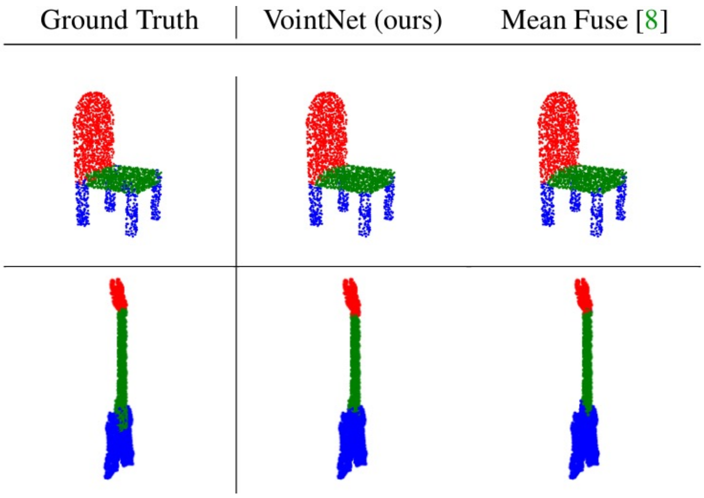


Here we compare our VointNet qualtltoavely to Mean fuse basline using the same pretrained 2D DeepLab V3 backbone and the GT



Note how we can find the details with VointNET that mean fuse misss like the window of the car

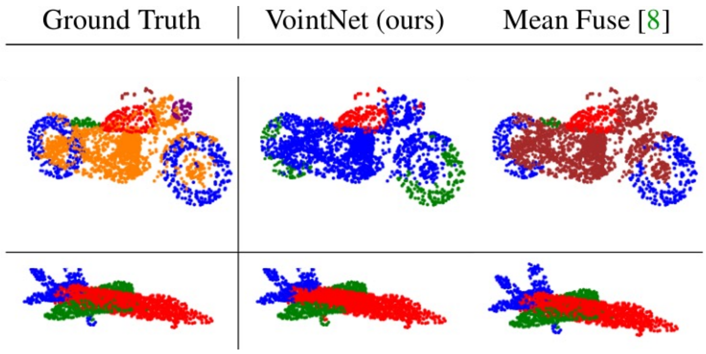
Qualitative Comparison

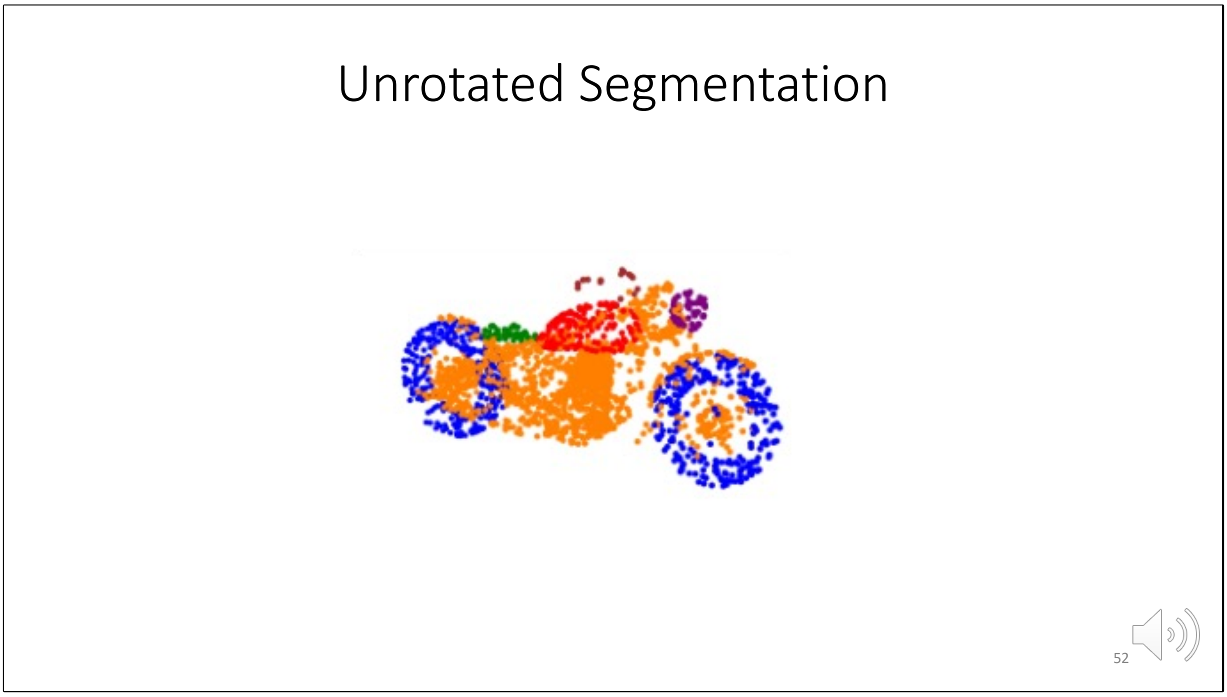


50

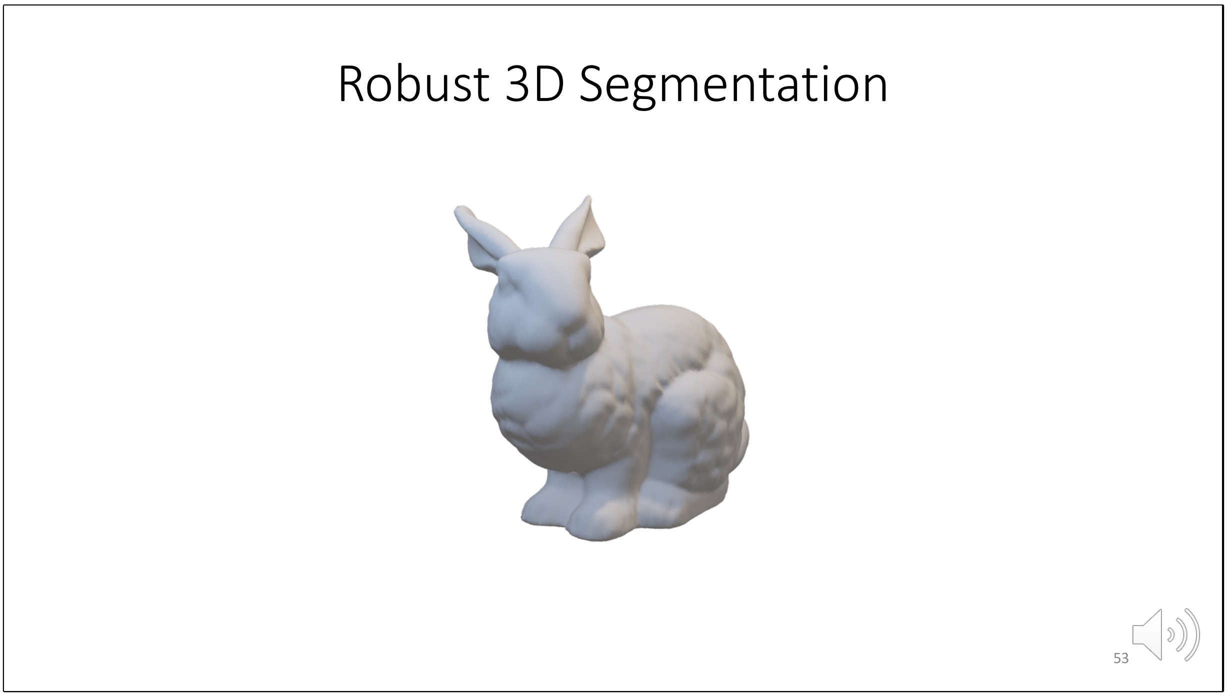
More comparisons

Qualitative Comparison





We also evaluate the robustness of our VointNet approach to rotaion by randomly rotating the object in test time in So(3).



We also evaluate the robustness of our VointNet approach to rotaion by randomly rotating the object in test time in So(3).

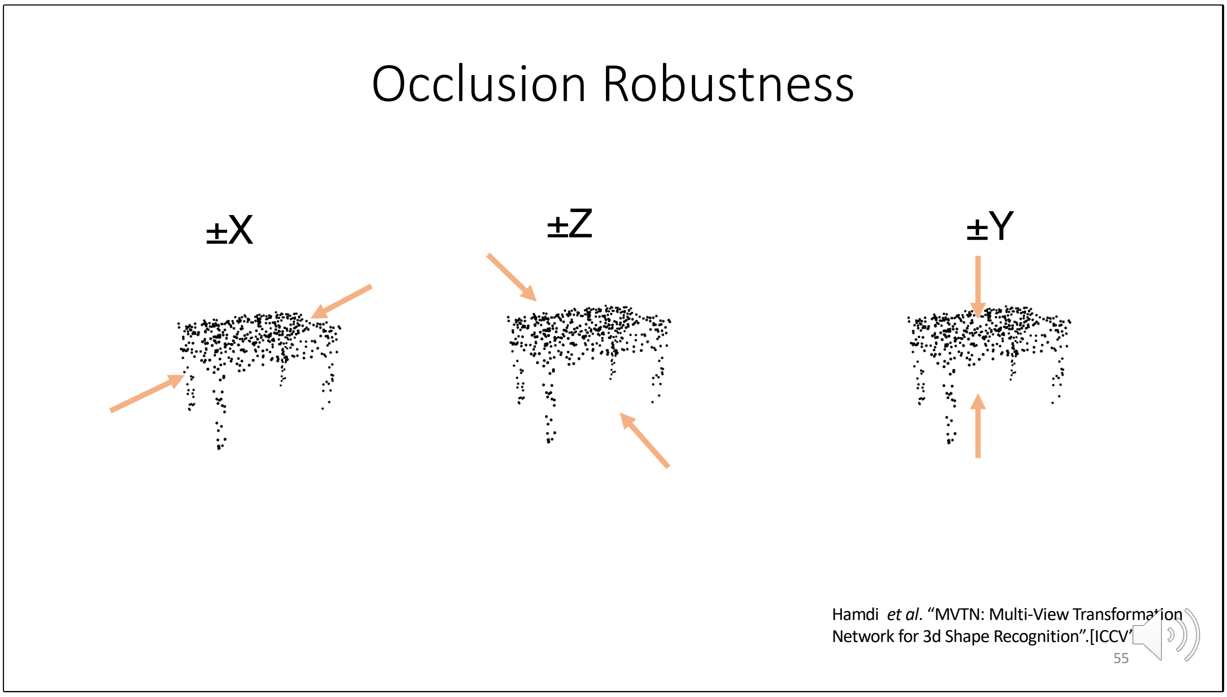
Robust 3D Part Segmentation

ShapeNet Parts

Method	Data Type	Part Segmentation (Unrotated)	(Rotated)
PointNet (Qi et al., 2017a)	Points	80.1	36.6 ±0.2
DGCNN (Wang et al., 2019c)	Points	80.1	37.1 ±0.2
CurveNet (Xiang et al., 2021)	Points	**84.9**	32.3 ±0.0
Label Fuse (Wang et al., 2019a)	M-View	80.0	61.4 ±0.2
Mean Fuse (Kundu et al., 2020)	M-View	77.5	62.0 ±0.2
VointNet (ours)	Voints	81.2	**62.4** ±0.2

54

ON ShapeNet Parts, we ahicve strong semgnetation perofmance on the aligned setup compared to other multi-view methods.and robust performnce tp rotation compared to point baselines



To simulate occlusion, we crop the object from its 6 faces with different percentages (0%-75%) and from different direwctions as in MVTN

Occlusion Robustness

Method	Data Type	Occlusion Ratio 0	0.1	0.2	0.3	0.5
PointNet (Qi et al., 2017a)	Points	89.1	88.2	86.1	81.6	53.5
DGCNN (Wang et al., 2019c)	Points	92.1	77.1	74.5	71.2	30.1
PCT (Guo et al., 2021)	Points	93.3	**92.6**	91.1	88.2	61.9
MVTN (Hamdi et al., 2021)	M-View	**93.8**	90.3	89.9	88.3	**67.1**
VointNet (ours)	Voints	92.8	91.6	**91.2**	**89.1**	66.1

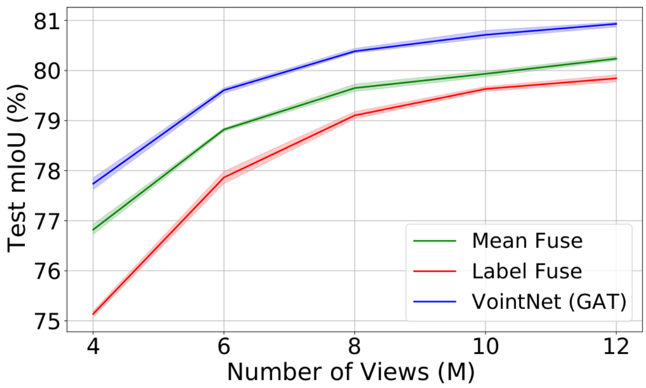
56

Here we show the average test accuracy on ModelNEt40 over the 6 canonical occlusion directions (± X, ± Y, ± Z) for different occlusion rations. VointNEt achove more robustness

Analysis

57

Number of View M



58

Here we study the effect of the number of views on the segmentation mIou performance for VointNet (Graph attention) , Mean fuse , and label fuse. All of the three use the exact same 2D backbone trained to segment the 2d projections of

GitHub Repo

https://github.com/ajhamdi/vointcloud

59

Code is Attached and will be made public

Thank You !

Please Check the paper and code for more details on ajhamdi/MVTN in github