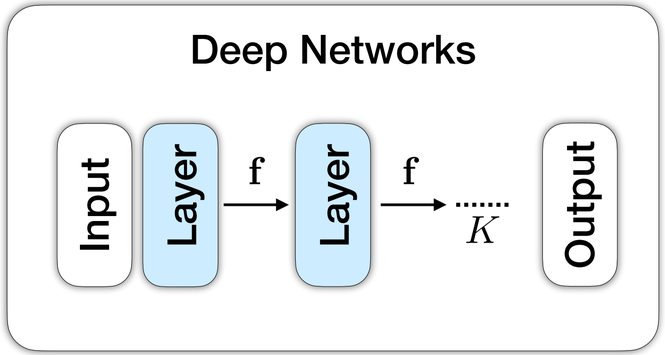
Exploring explainable models for point cloud learning

Weiwei Sun

5th Ph.D. student @ UBC

under Dr. Andrea Tagliasacchi & Dr. Kwang Moo Yi

Explainability is essential



A stack of *K layers* and a set of intermediate features f.

As a black box, deep networks work pretty good

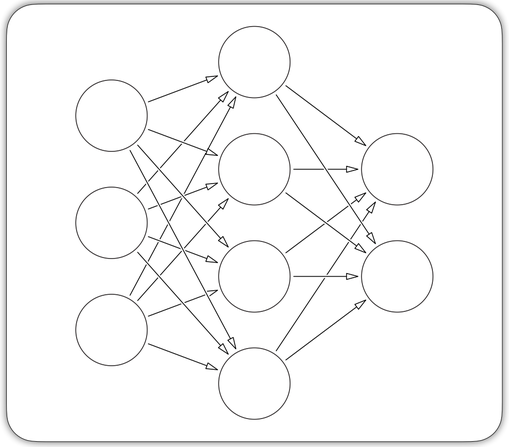
but lacks transparency, controllability, generalization ability.

How does deep networks behave as human understand? And, to what extent, we can benefit from explainable models?

Point cloud learning

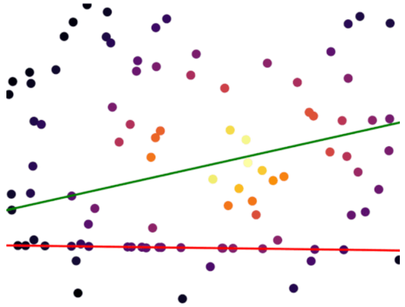
Permutation-equivariant point network

$$\mathbf{P} = \begin{Bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_N \end{Bmatrix}$$



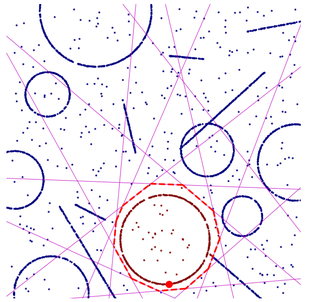
$$\mathbf{F} = \begin{Bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_N \end{Bmatrix}$$

We present models explainable in either layer or features.



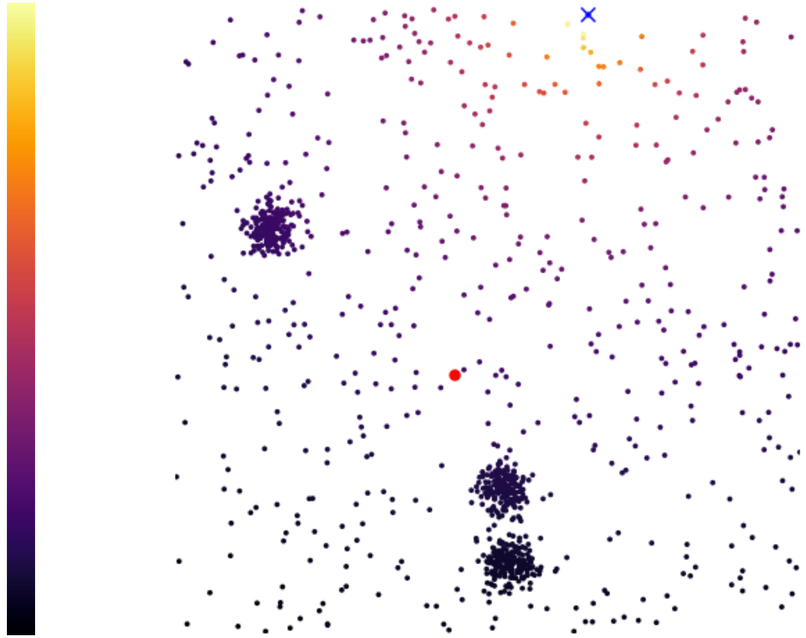
ACNe@CVPR2020

Canonical Capsules@NeurIPS2021



NeuralBF@WACV2023

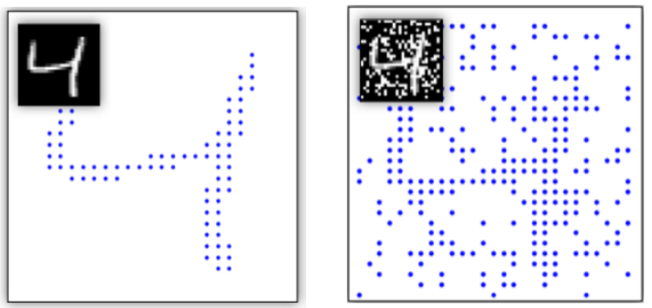
ACNe: unrolling optimization as deep network

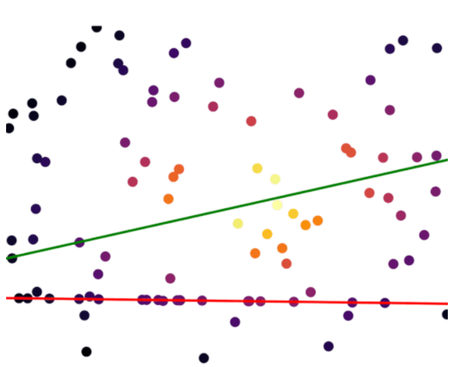


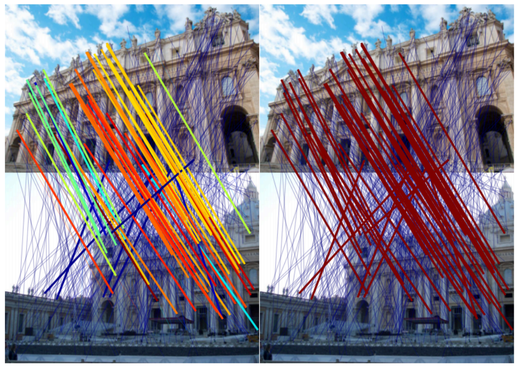
● Mean

× **Weighted mean**

Inspired by IRLS — treating each step as a layer.







Matches: ACN Matches: Ground Truth

SOTA performance in robust point cloud learning tasks.

Sun, Jiang, Trulls, Tagliasacchi, Yi «ACNe: attentive context normalization for robust permutation-equivariant learning» CVPR 2020

Robust Learning of Point Clouds

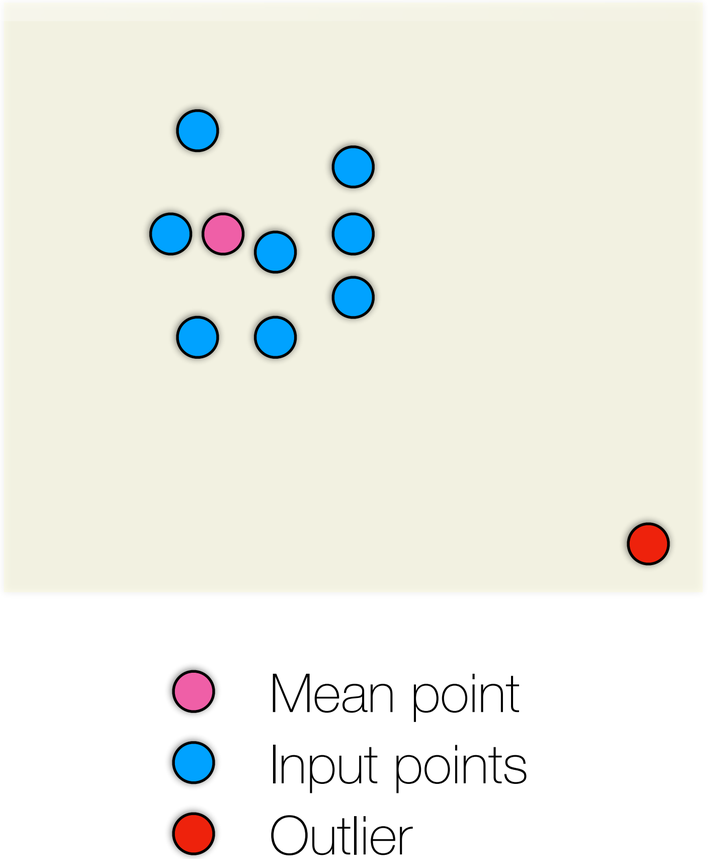
Feature normalization relates points

- Context normalization

$$\mathcal{N}_{CN}(\mathbf{f}) = (\mathbf{f} - \mu(\mathbf{f})) \oslash \sigma(\mathbf{f})$$

- Where mean and average can be rewritten as

$$\mu(\mathbf{f}) = \mathbb{E}[\mathbf{f}]$$
$$\sigma(\mathbf{f}) = \sqrt{\mathbb{E}[(\mathbf{f} - \mathbb{E}[\mathbf{f}])^{\circ 2}]}$$



Sun, Jiang, Trulls, Tagliasacchi, Yi «ACNe: attentive context normalization for robust permutation-equivariant learning» CVPR 2020

Robust Learning of Point Clouds

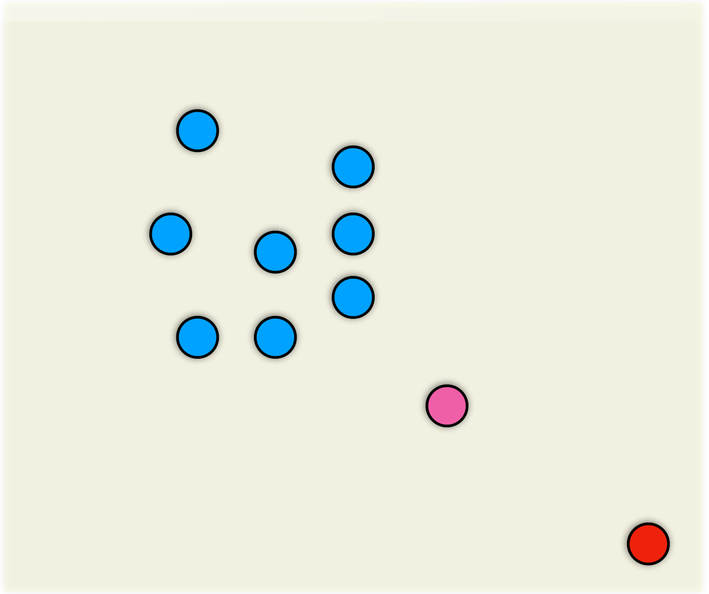
Attentive Context Normalization (ACN)

- So we define our attention based on features

$$\mathbf{w} = \eta(\mathcal{W}_{\boldsymbol{\omega}}(\mathbf{f}))\,, \quad \eta(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_1.$$

- And then define an attentive normalization

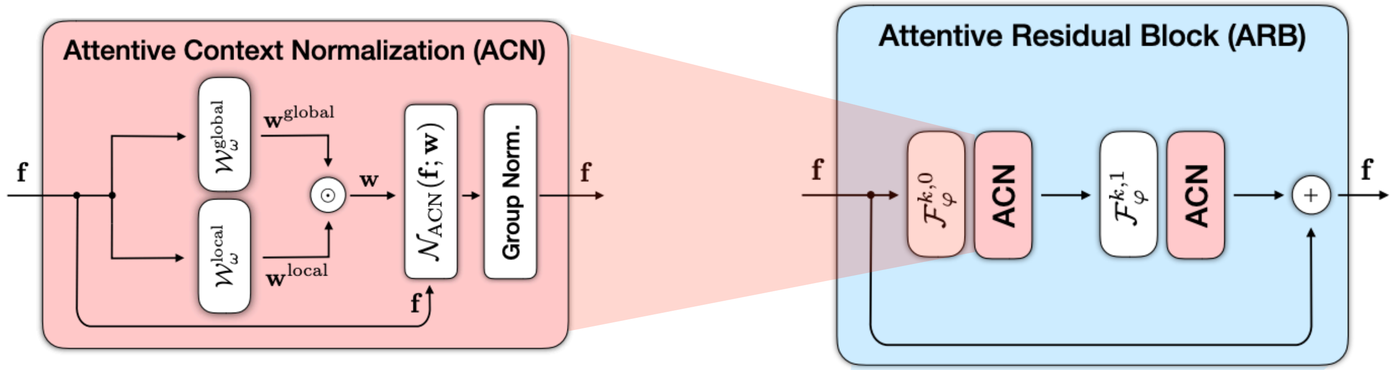
$$\mathcal{N}_{ACN}(\mathbf{f};\mathbf{w}) = (\mathbf{f} - \mu_{\mathbf{w}}(\mathbf{f})) \oslash \sigma_{\mathbf{w}}(\mathbf{f}).$$



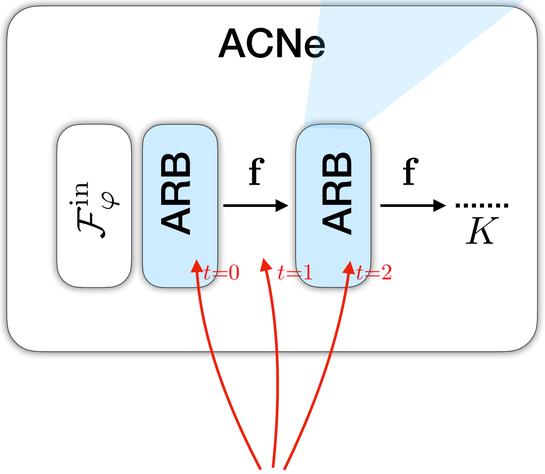
Mean point
Input points
Outlier

Sun, Jiang, Trulls, Tagliasacchi, Yi «ACNe: attentive context normalization for robust permutation-equivariant learning» CVPR 2020

ACNe: an iterative network



$$\mathbf{P} = \begin{Bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_N \end{Bmatrix}$$

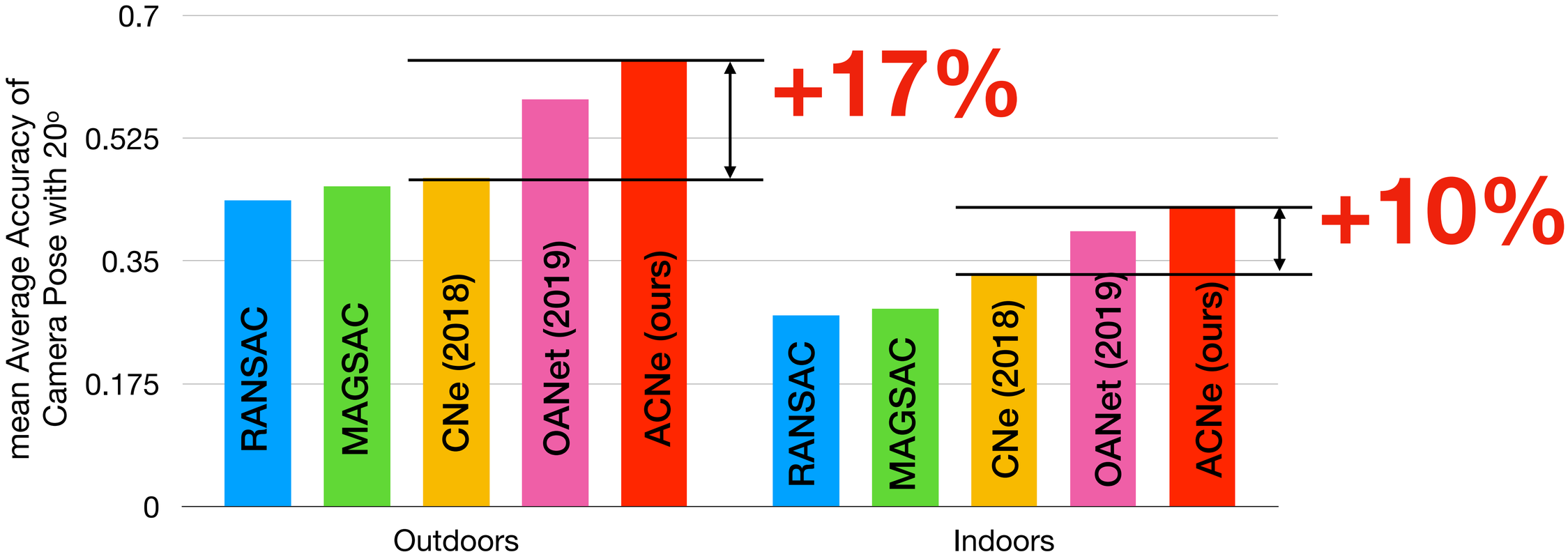


$$\mathbf{F} = \begin{Bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_N \end{Bmatrix}$$

Sun, Jiang, Trulls, Tagliasacchi, Yi «ACNe: attentive context normalization for robust permutation-equivariant learning» CVPR 2020

ACNe: explainable model won with less paramters

In real-world wide-baseline stereo matching task, explainable ACNe achieves the best performance yet with less parameters.

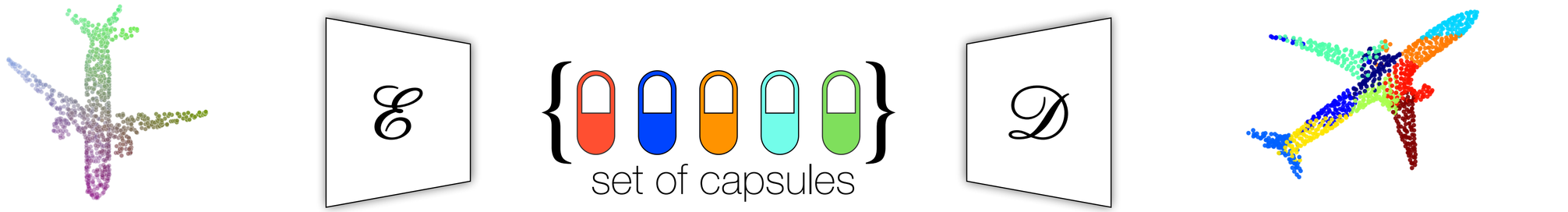


Methods	OANet	CNe	ACNe
# of Parameters	2347K	394K	400K

Sun, Jiang, Trulls, Tagliasacchi, Yi «ACNe: attentive context normalization for robust permutation-equivariant learning» CVPR 2020

Canonical Capsules: explainable features

- A self-supervised training architecture for point clouds with:
  - SOTA performance in auto-encoding
  - SOTA performance in unsupervised classification
  - SOTA performance in learnt canonicalization



input point cloud
(randomly transformed)

auto-encoding
(learnt canonical frame)

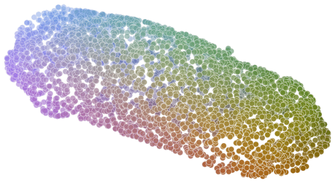
Sun*, Tagliasacchi*, Deng, Sabour, Yazdani, Hinton, Yi «Canonical Capsules: Self-Supervised Capsules in Canonical Pose» NeurIPS 2021

Built-in biases in 3D datasets

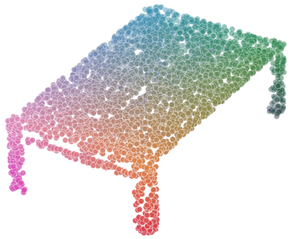
chairs

airplane

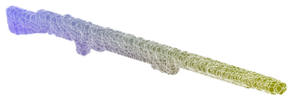
car

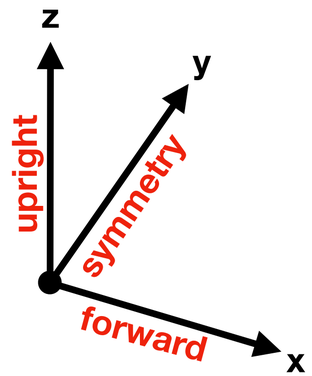


table



firearm

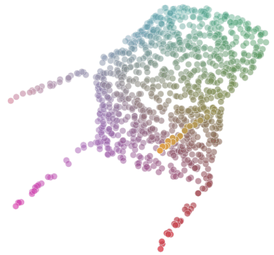




Chang et al. arXiv 2015 «ShapeNet: An Information-Rich 3D Model Repository»

Auto-encoding Point Clouds

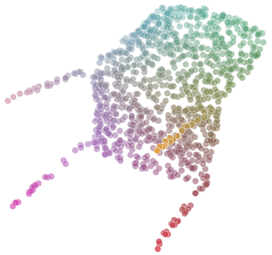
ShapeNet
aligned data



AtlasNetV2

performance
drop

ShapeNet
non-aligned data



AtlasNetV2

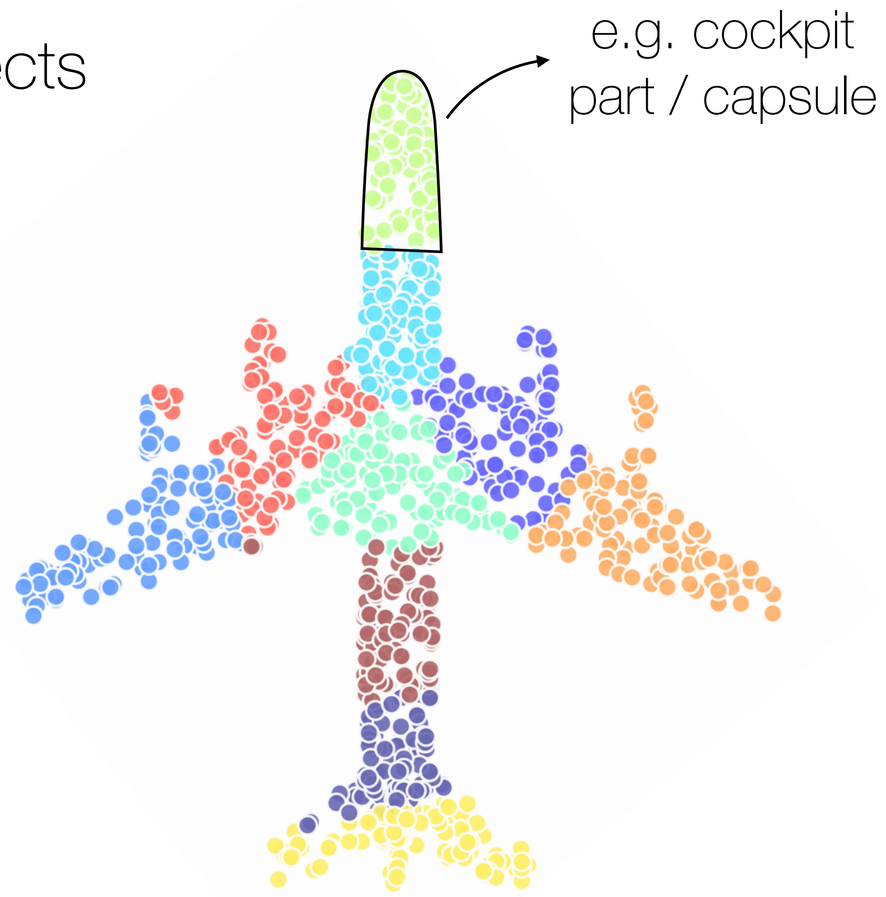
Deprelle et al. NeurIPS 2019 «AtlasNetV2: Learning elementary structures for 3D shape generation and matching"

But… what are capsules?

- An abstraction of scene graphs from computer graphics
- Capsules are part-based representations of objects
- Think of a capsule  as an “oriented part”

capsule descriptor
part shape

capsule pose
part orientation

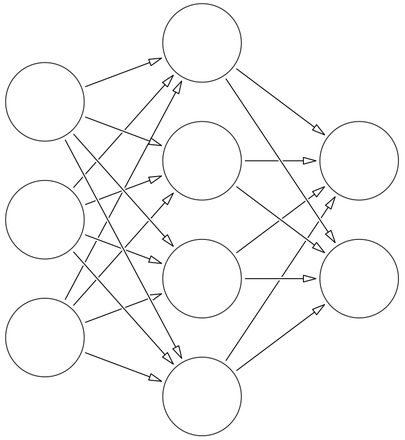


Credit to: Hinton et al. NeurIPS 2011 «Transforming Auto-encoders»

Sun*, Tagliasacchi*, Deng, Sabour, Yazdani, Hinton, Yi «Canonical Capsules: Self-Supervised Capsules in Canonical Pose» NeurIPS 2021

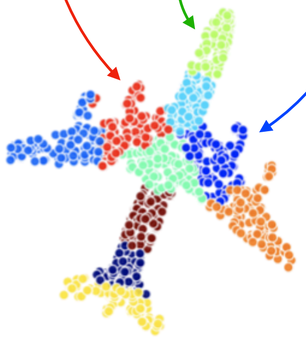
Multi-head attention (ACNe)

$$\mathbf{P} = \begin{Bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_N \end{Bmatrix}$$



ACNe

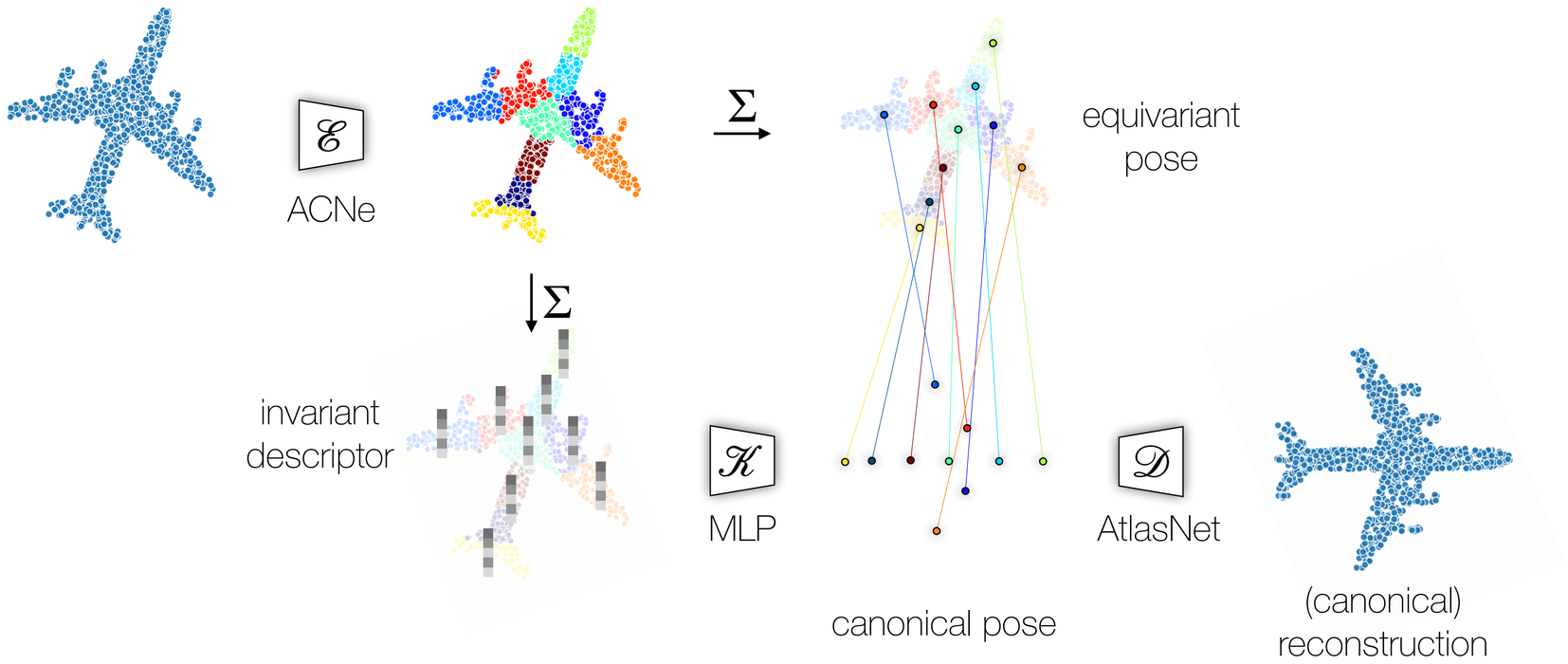
$$\begin{Bmatrix} w_1^1 & w_1^2 & w_1^3 & \dots & w_1^k \\ w_2^1 & w_2^2 & w_2^3 & \dots & w_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_N^1 & w_N^2 & w_N^3 & \dots & w_N^k \end{Bmatrix} = \mathbf{A}$$



Sun*, Tagliasacchi*, Deng, Sabour, Yazdani, Hinton, Yi «Canonical Capsules: Self-Supervised Capsules in Canonical Pose» NeurIPS 2021

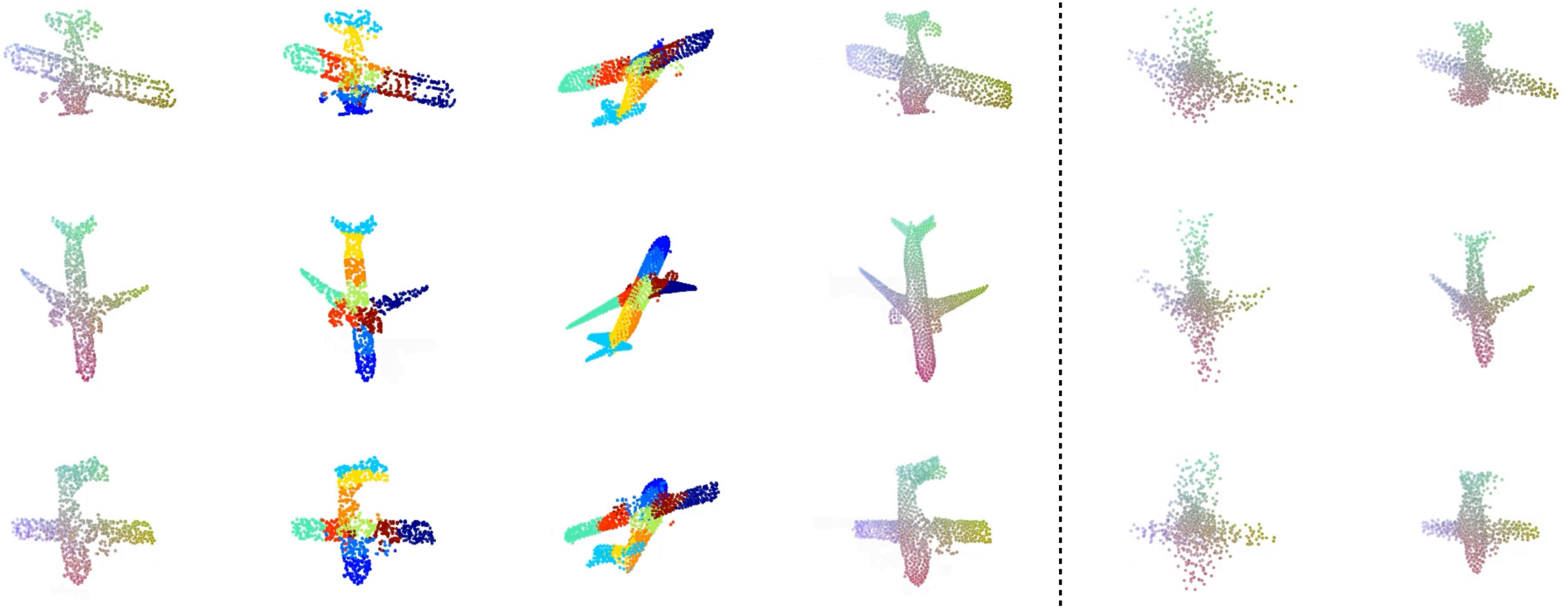
Inference Procedure

- Let us look at how the inference (forward pass) of our network architecture
  - Note the input is a point cloud in any pose (random rotation/translation)
  - While the reconstruction is executed in the canonical frame



Sun*, Tagliasacchi*, Deng, Sabour, Yazdani, Hinton, Yi «Canonical Capsules: Self-Supervised Capsules in Canonical Pose» NeurIPS 2021

Auto Encoding – Controllability and better performance



Input (Point Cloud)

Decomposition (original frame)

Reconstruction (canonical frame)

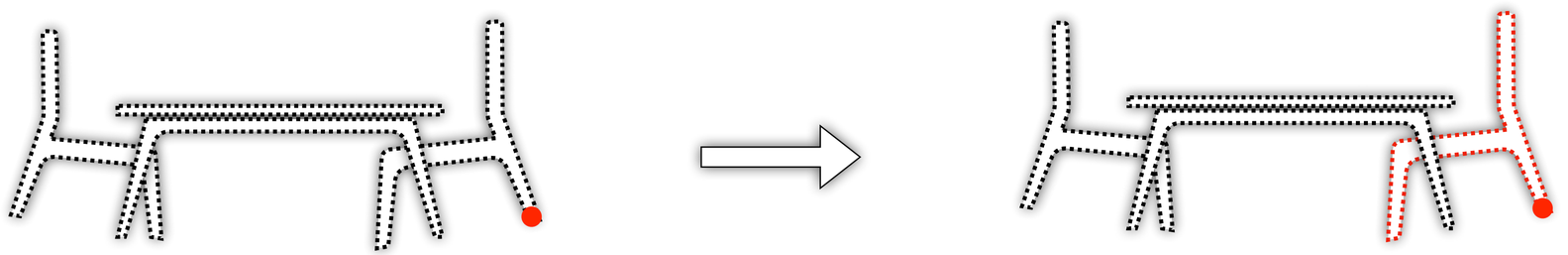
Reconstruction (original frame)

PointCapsNet (CVPR 2019)

AtlasNetV2 (NeurIPS 2019)

NeuralBF: Explainable feature and layer

- An iterative layer that
  - is inspired by traiditional bilateral filtering
  - generates query-conditioned instance proposals for 3D point clouds
  - gives state-of-the-art top-down 3D instance segmentation.



Query

Affinity

Sun, Rebain, Liao, Tankovich, Yazdani, Yi, Tagliasacchi «NeuralBF: Neural Bilateral Filtering for Top-down Instance Segmentation on Point Clouds» WACV2023

Direct bounding box regression

- A commonly used method for instance proposal generation
- Widely adopted for top-down 2D instance segmentation
- However, bottleneck the performance of top-down 3D instance segmentation

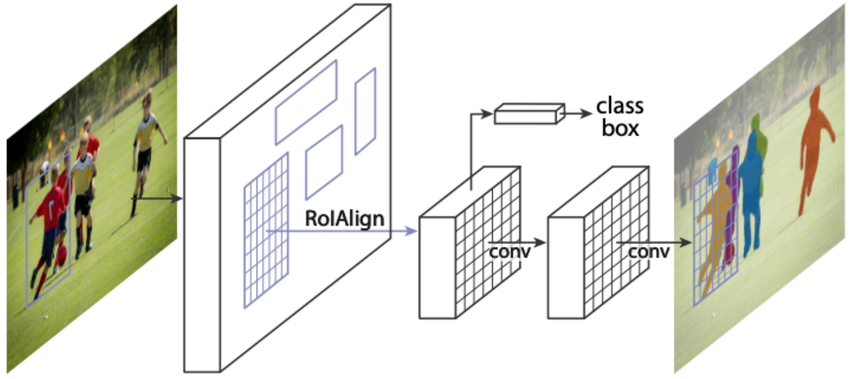
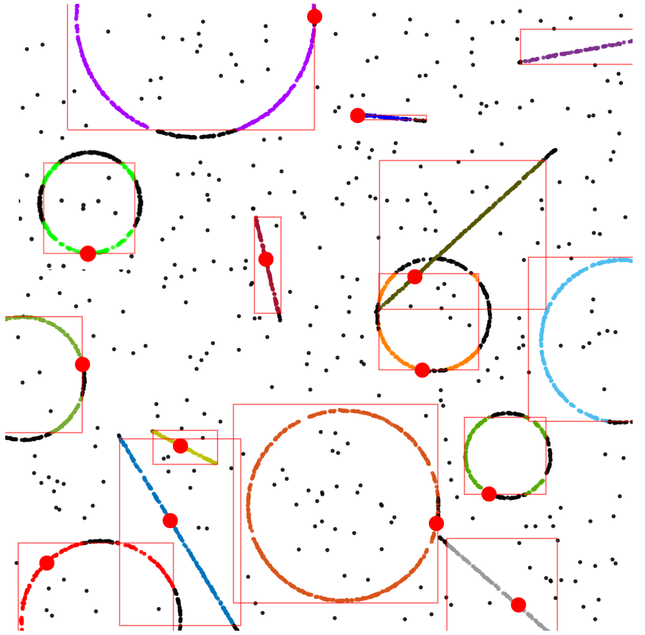


Figure 1. The **Mask R-CNN** framework for instance segmentation.

He et al. Mask R-CNN. ICCV 17.

Great success in 2D images

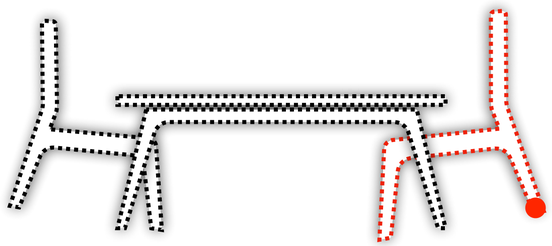
To 3D point cloud



mAP < 50%

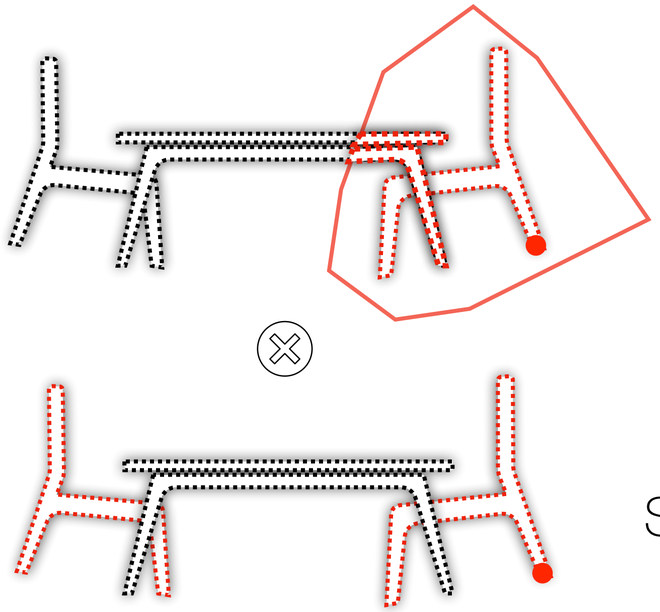
Affinity as proposal

- Input a given query (q) on the input point cloud.
- Output affinity score map of range [0, 1].
- Consists of *spatial* and *semantic* affinity.



Affinity

$\mathcal{A}(q)$



$\mathcal{A}_{\mathbf{p}}(q)$

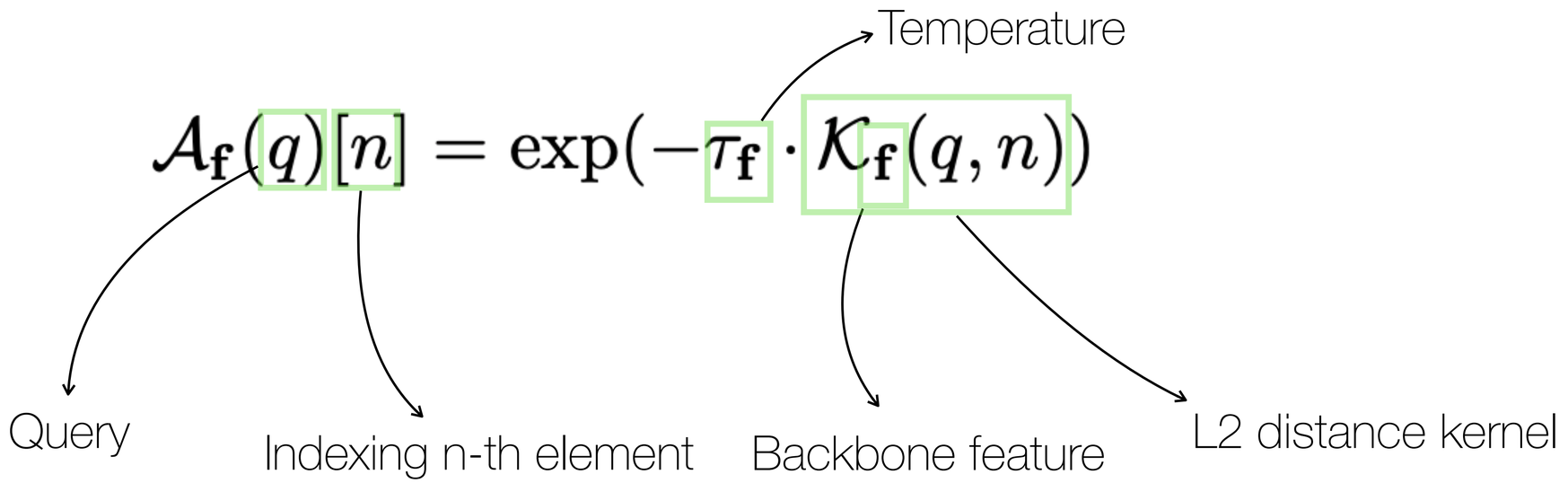
Spatial similarity

Semantic similarity

$\mathcal{A}_{\mathbf{f}}(q)$

Semantic similarity

- A simple exponential kernel
  - L2 distance between semantic similarity features

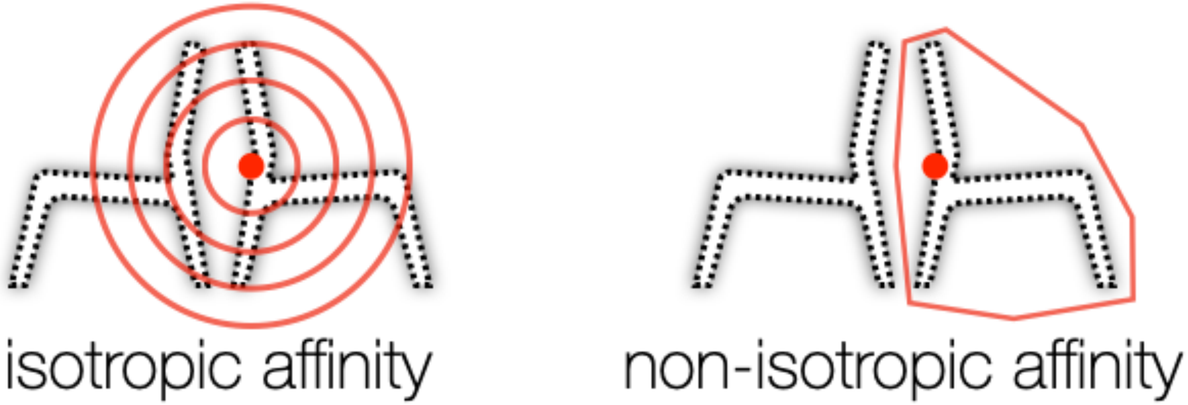


Spatial similarity

- A non-isotropic kernel based on local convex hulls
  - Is able to isolate the queried instance

$$\mathcal{A}_{\mathbf{p}}(q)[n] = \exp(-\tau_{\mathbf{p}} \cdot \mathcal{K}_{\mathbf{p}}(q, n))$$

Temperature

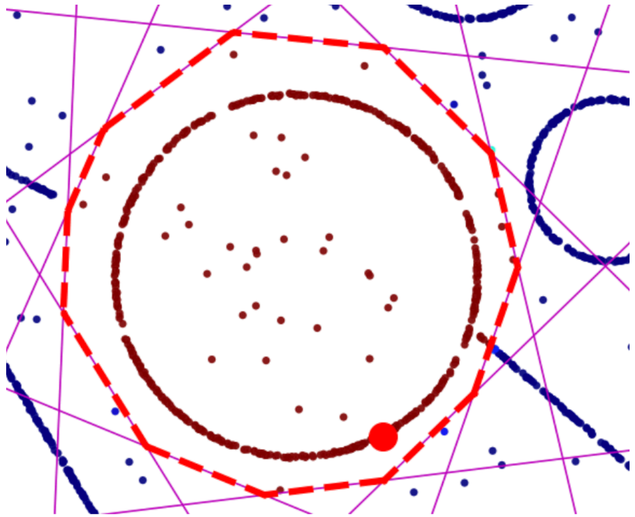


isotropic affinity

non-isotropic affinity

Spatial similarity — in more details

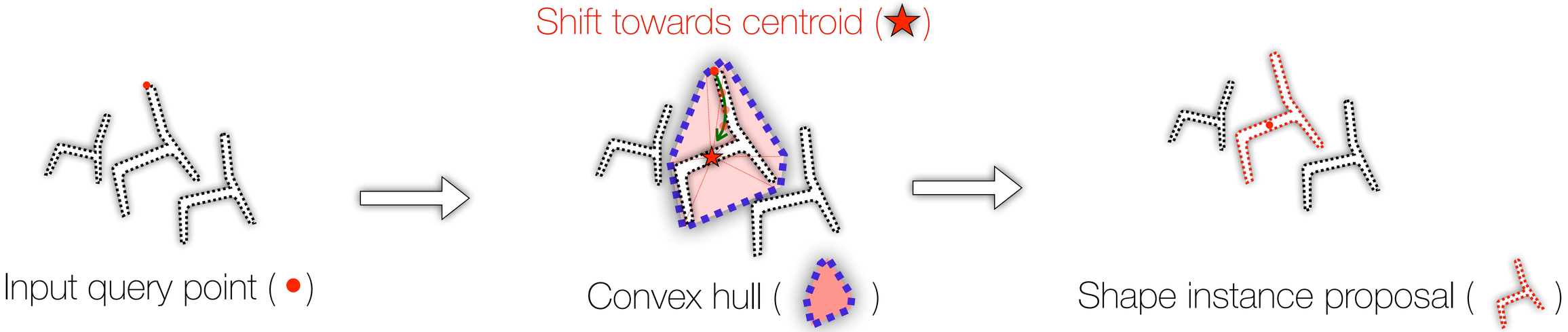
- A hyper coordinate network -- convex hull
  - Consists of a set of hyperplanes conditional on query
  - Is an explainable neural field—comparing with recent MLP-based neural field.
  - Is very small and memory efficient



$$\mathcal{C}(\mathbf{x};\mathbf{f}) \begin{cases} = 0 & \text{if } \mathbf{x} \text{ inside convex defined by } \mathbf{f}, \\ > 0 & \text{otherwise } (\approx \text{boundary distance}). \end{cases}$$

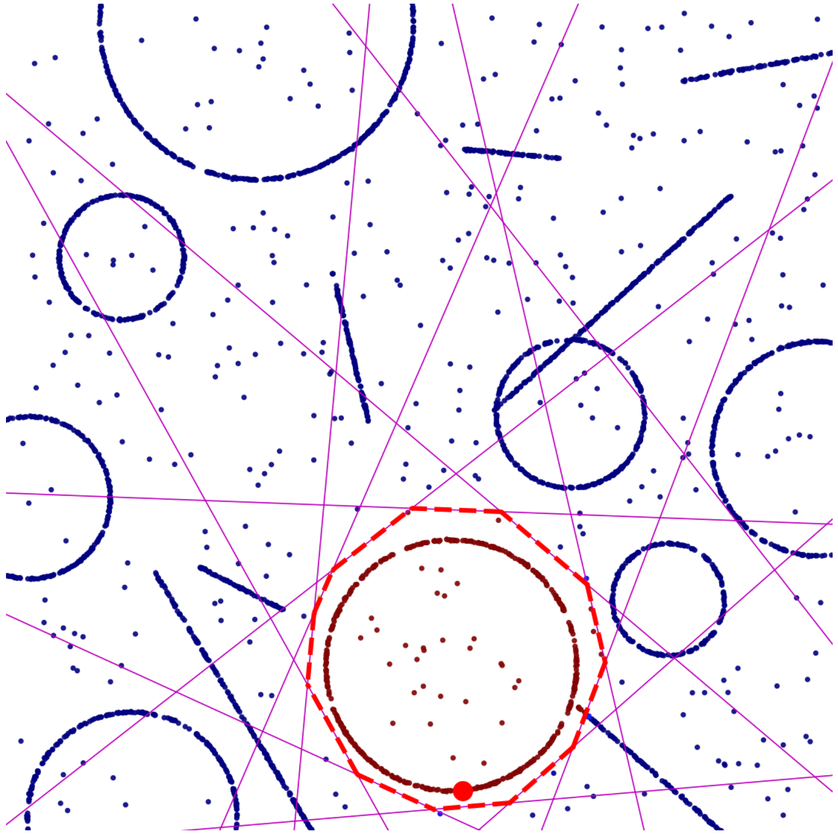
Bilateral filtering

- Iterative inference
  - Update query as the weighted center of input
  - Shift query into the centroid of instance

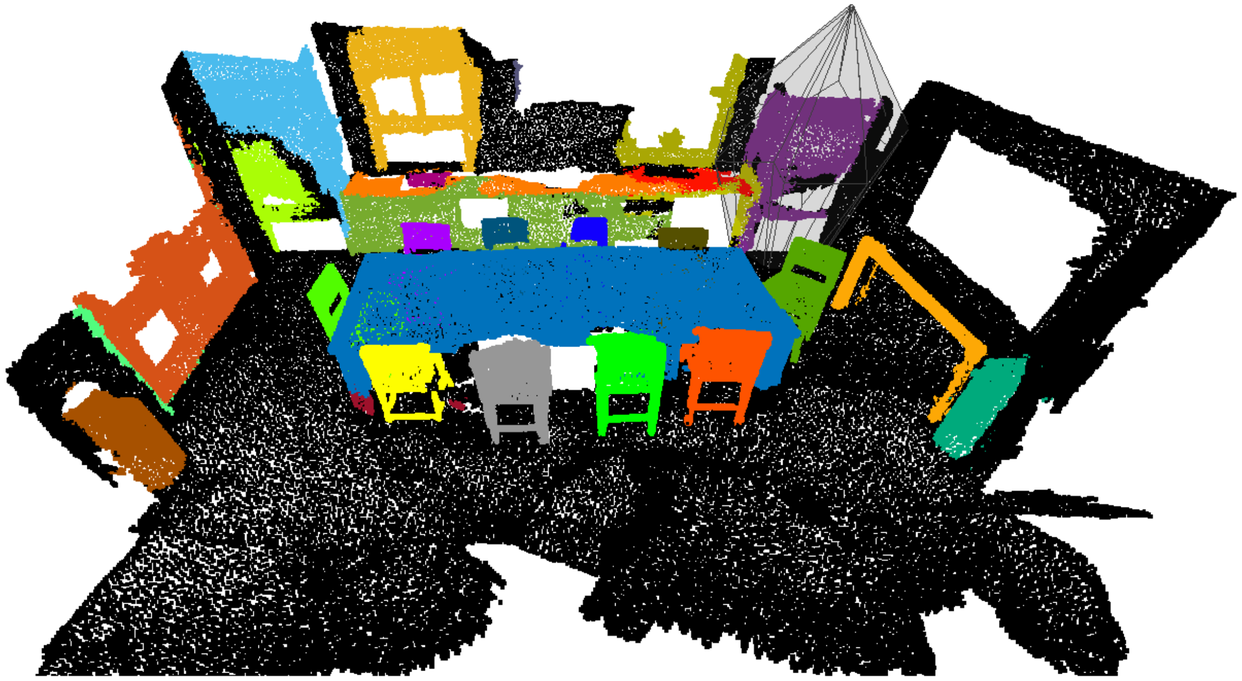


Results — Learned convex hull of high quality

Convex hulls in 2D



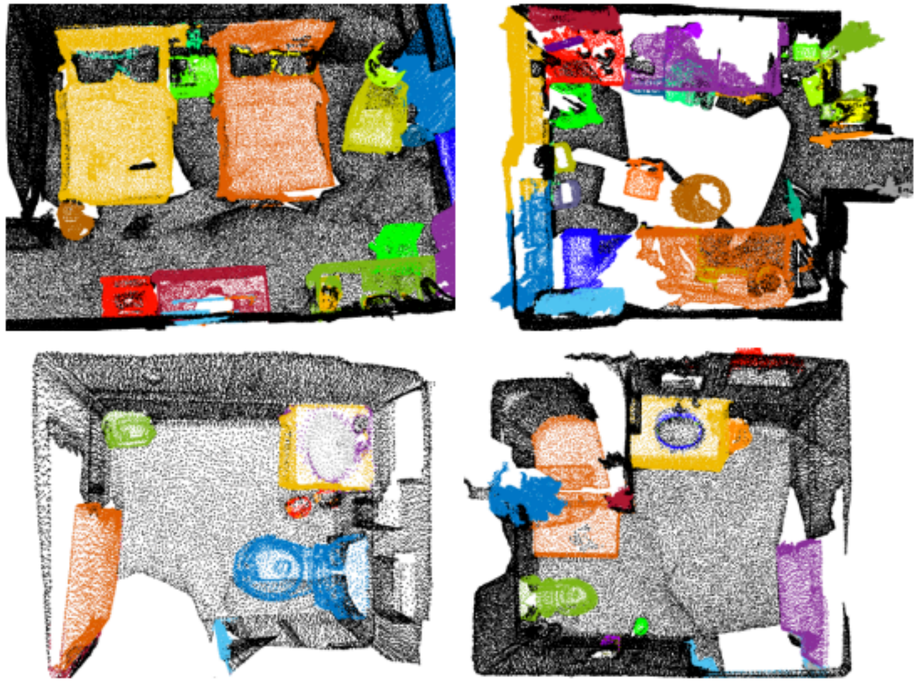
Convex hulls in 3D



Results — Instance segmentation in ScanNet.

- Our method is the best among purely top-down methods
- Top-down methods in general lags behind the bottom-up/Mix strategy

Qualitative



Quantitative

	Methods	Validation			Test		
		mAP	AP_{50}	AP_{25}	mAP	AP_{50}	AP_{25}
Bottom-up	PointGroup [22]	34.8	56.7	71.3	40.7	63.6	77.8
	SSTNet [26]	49.4	64.3	74.0	50.6	69.8	78.9
	HAIS [2]	43.5	64.4	75.6	45.7	69.9	80.3
Mix	Dyco3D [18]	35.4	57.6	72.9	39.5	64.1	76.1
	SoftGroup [44]	–	67.6	78.9	50.4	76.1	86.5
Top-down	3D-SIS [20]	–	18.7	35.7	16.1	38.2	55.8
	GSPN [53]	19.3	37.8	53.4	–	30.6	–
	3D-Bonet [52]	–	–	–	25.3	48.8	68.7
	Ours	**36.0**	**55.5**	**71.1**	**35.3**	**55.5**	**71.8**

Summary

Explainable models have lots of potential. When carefully designed, these models can also have good performance!

wsunid.github.io