CSC2457 3D & Geometric Deep Learning

On Learning Sets of Symmetric Elements
(Best paper award, ICML 2020)

Haggai Maron, Or Litany, Gal Chechik, Ethan Fetaya

Date: 03-09-2020

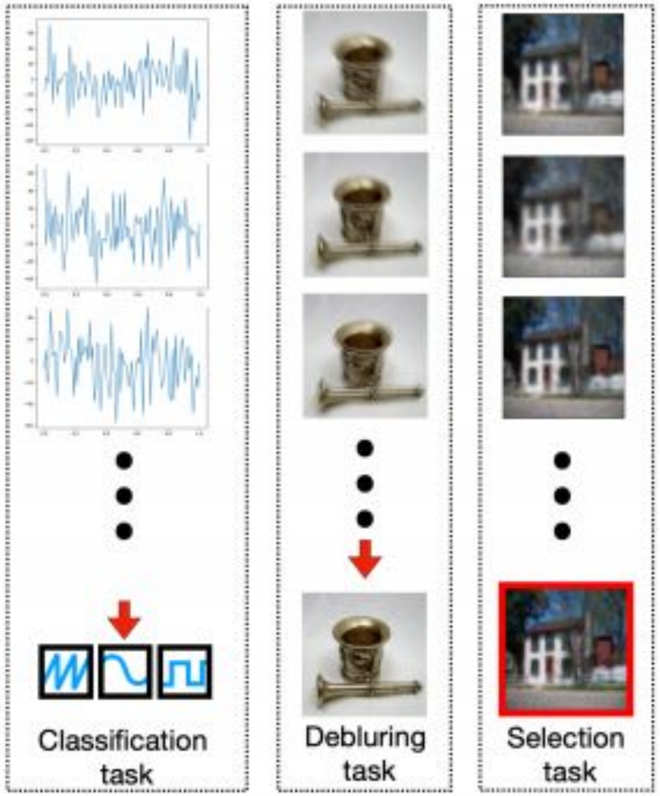
Presenter: Dmitrii Shubin

Instructor: Animesh Garg

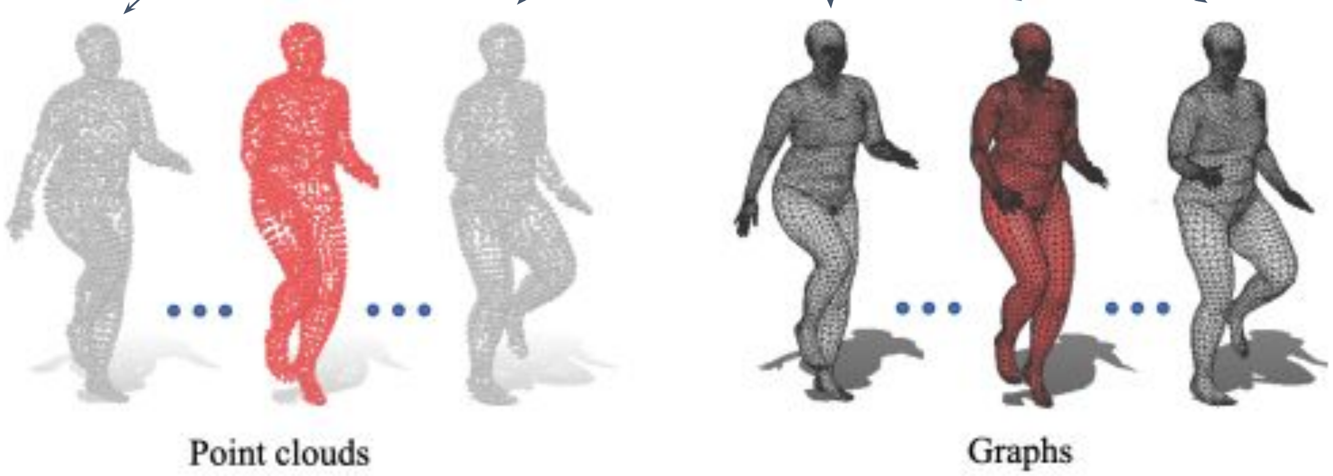
UNIVERSITY OF TORONTO

Motivation: Applications

We don't care about samples' position



Also don't care

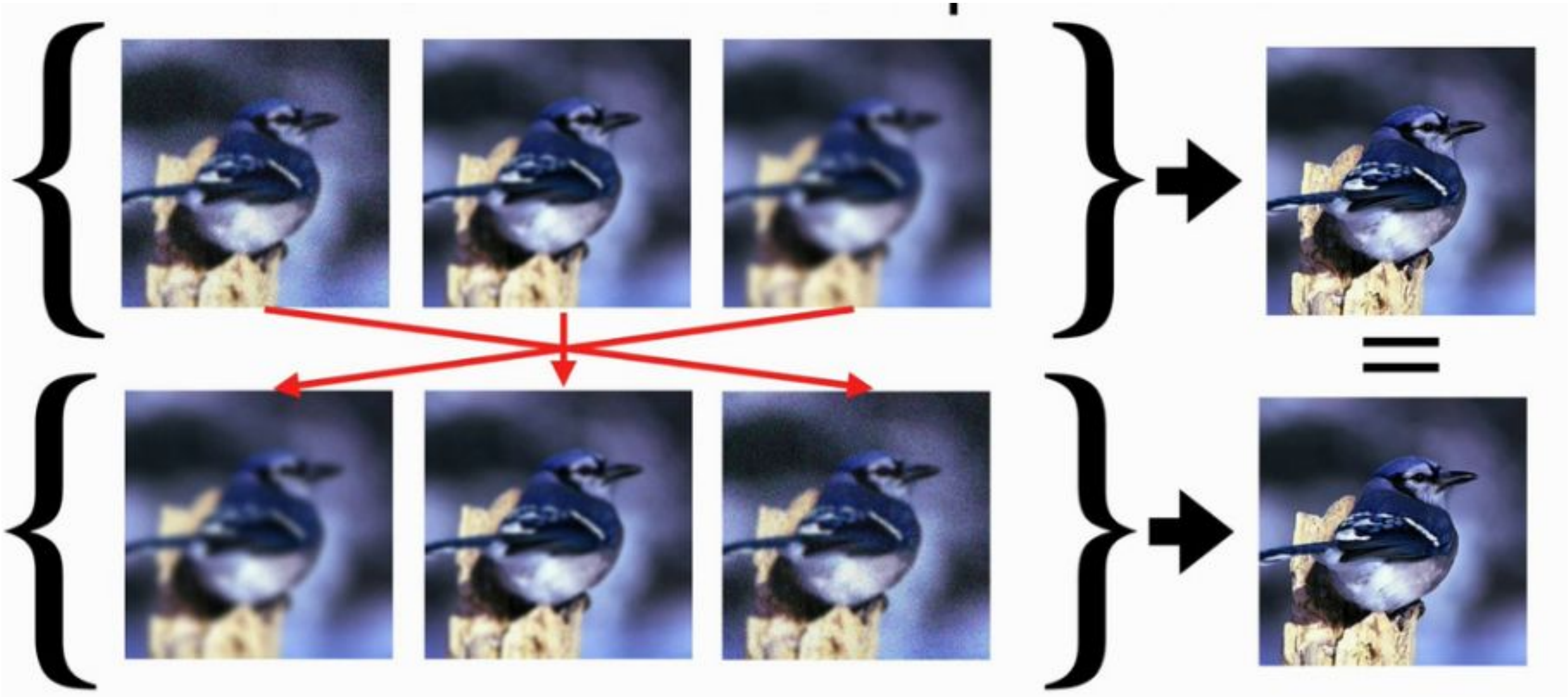


Frame selection tasks

Contributions

- Formulated the problem of limited expressive power of Deep Sets architectures
- Proposed a model architecture called DSS, both invariant and equivariant to the order/translation of input samples
- Mathematically proved of the proposed model architecture
- Provided a benchmarks on a variety sets of tasks and various data types

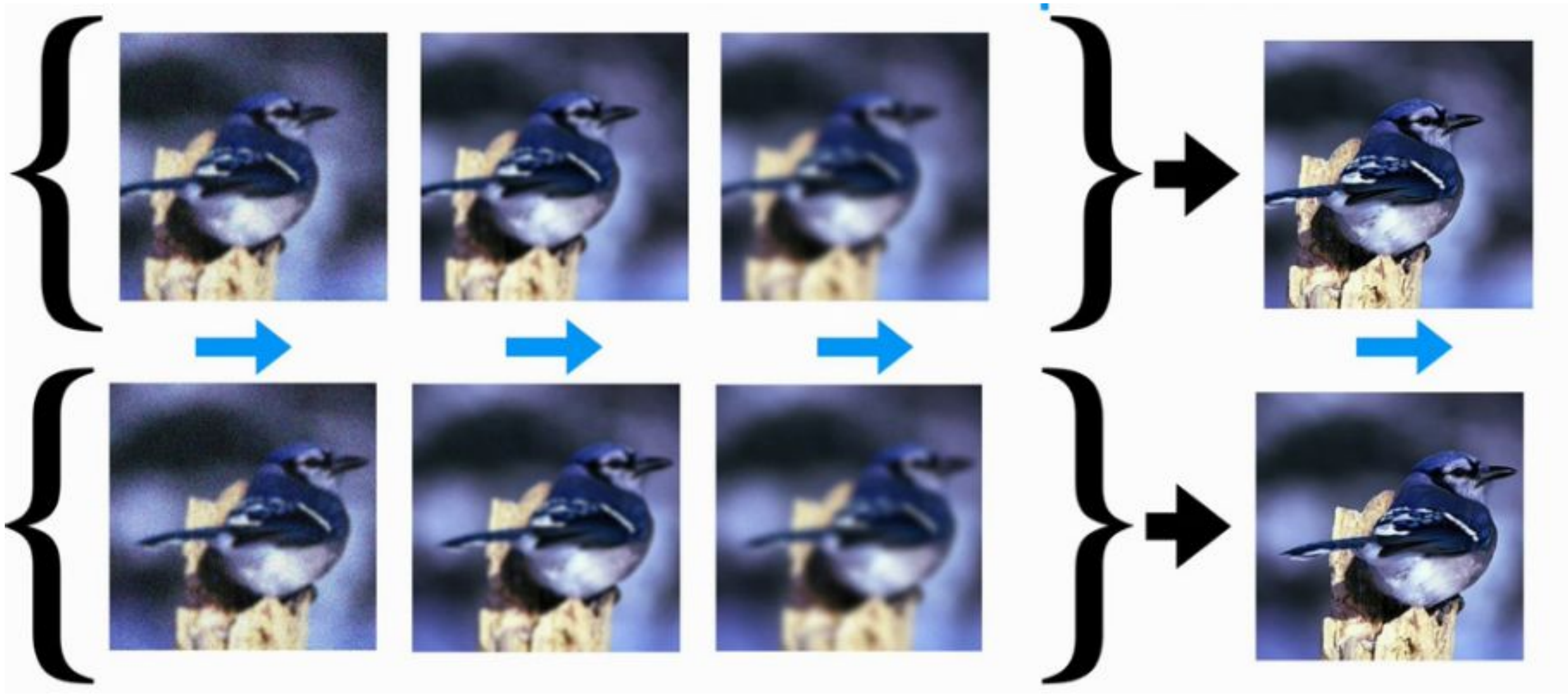
Recap: Translation invariance



=

Invariant to the order of the images in set

Recap: Translation equivariance



=

equivariant to the shifts of the images in set

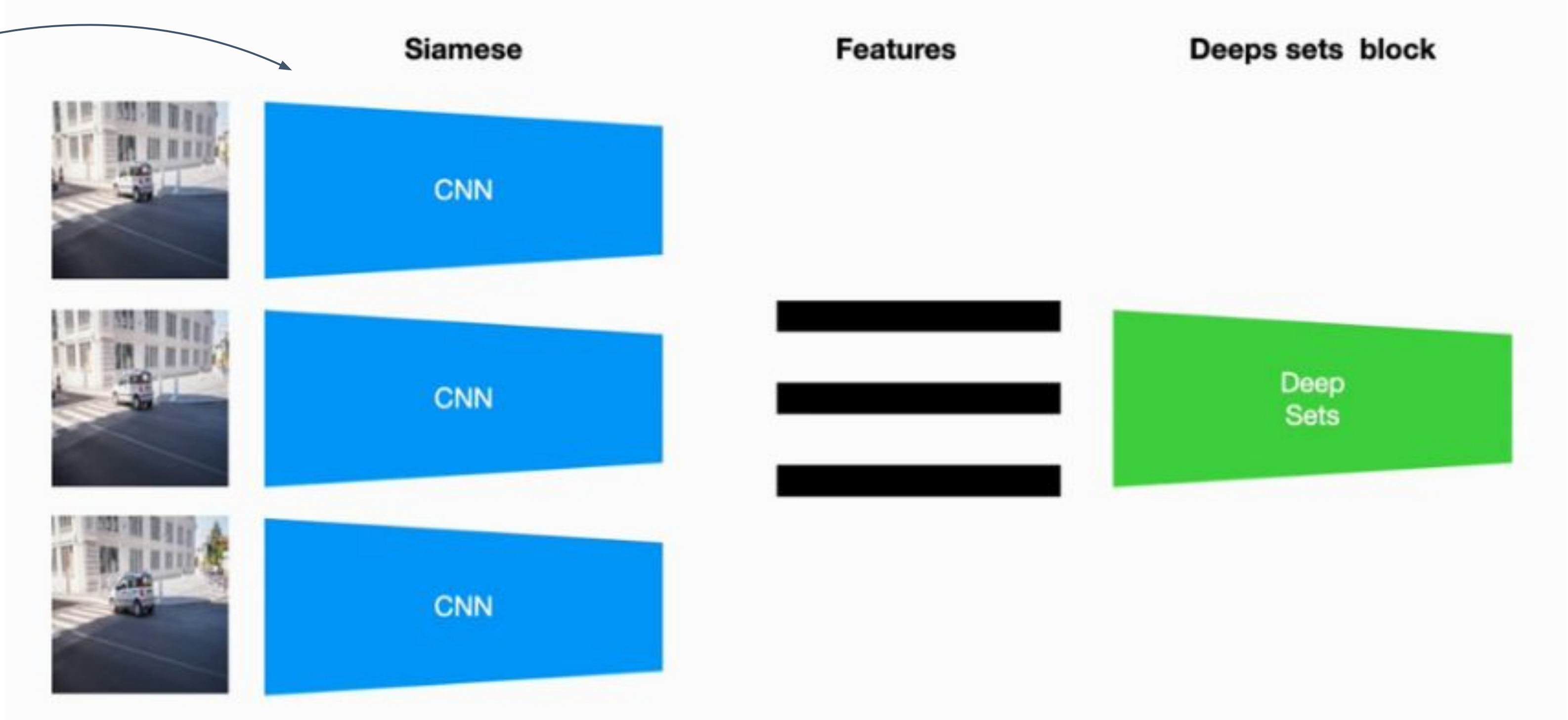
Recap: Deet Sets

No shared information between elements in the set

=

limited expressive power

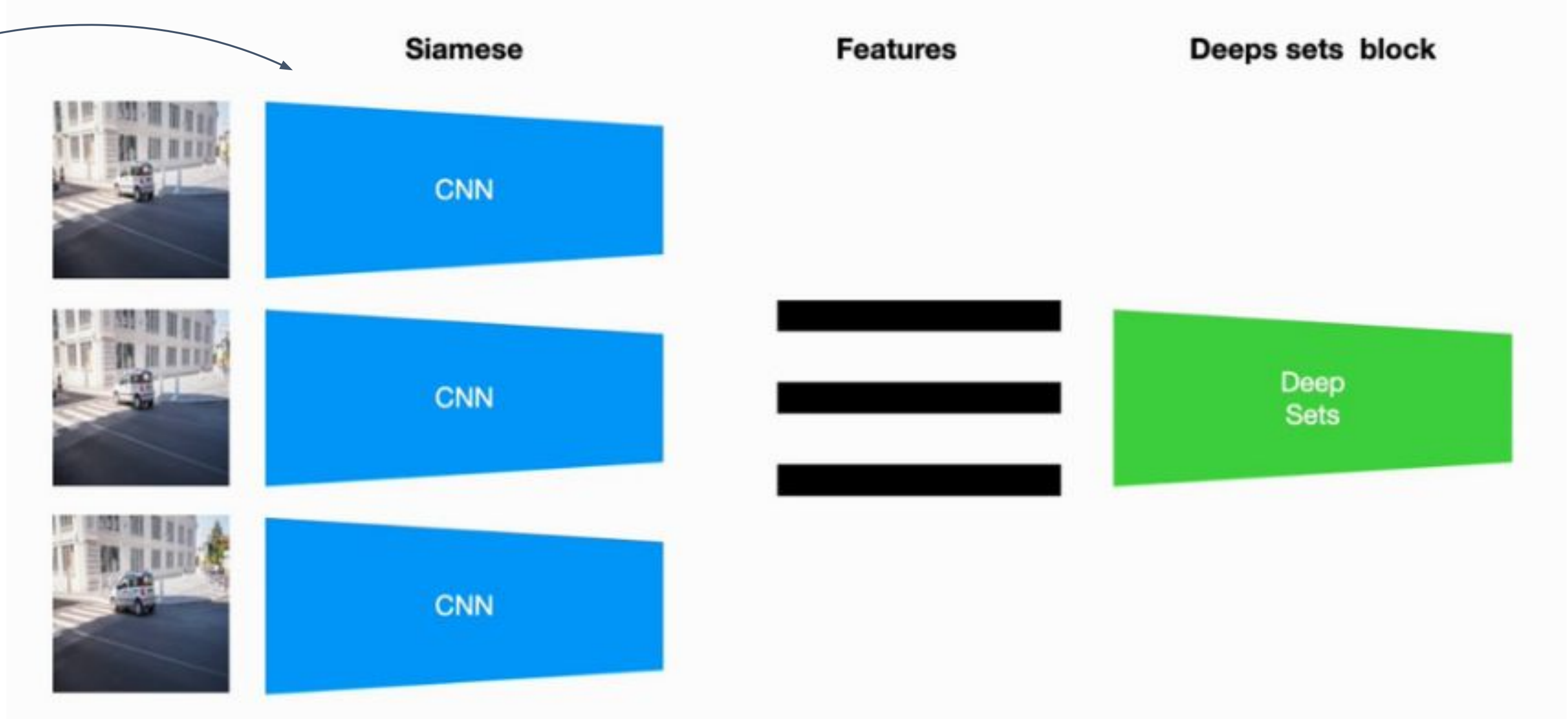
(i.e. degree of freedom)



Deep Sets, Zaheer et al., 2017

Recap: Deet Sets

"Late feature aggregation"



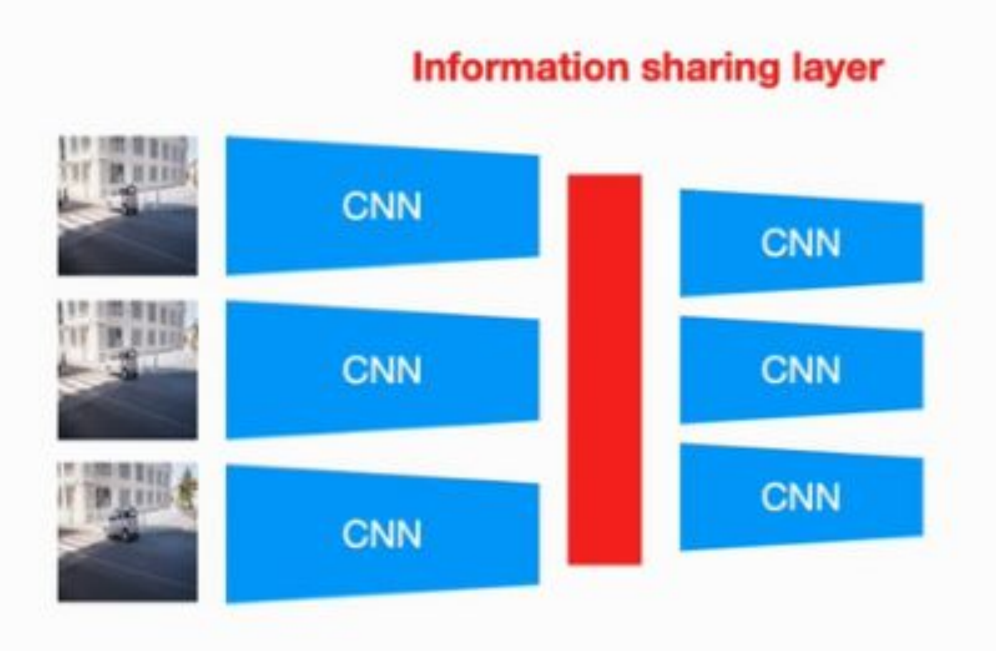
Deep Sets, Zaheer et al., 2017

Recap: Attempts to add information sharing

Aittala, Durand, ECCV 2018

Sridhar et al., NeuriPS 2019

Liu et al., ICCV 2019



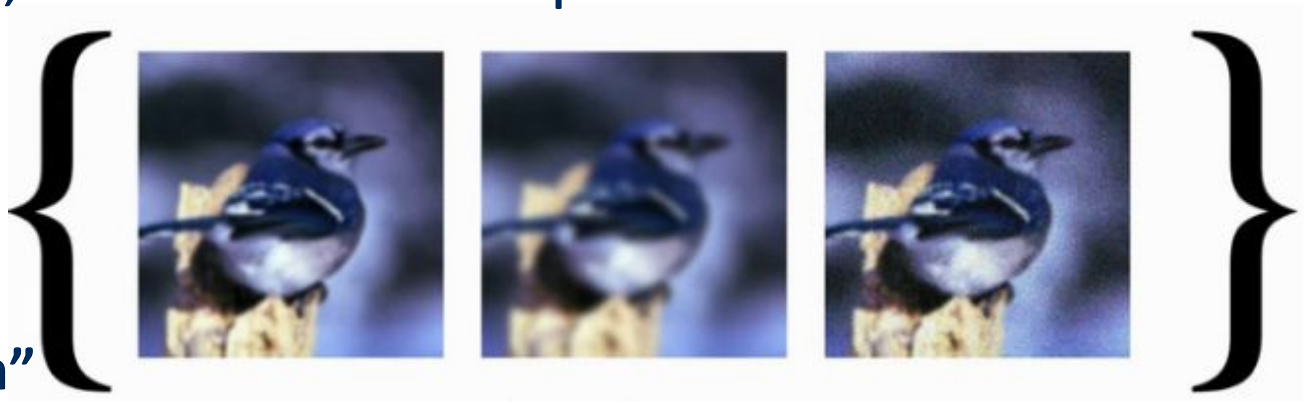
Problem Setting

How to create a such model architecture that both translate invariant and equivariant without loss of the expressive power and do it efficiently?

Problem Setting

Number of sets, i.e. translation equivariance

H



“H-equivariant/equivarian”

“G-invariant/equivariant”

1

n

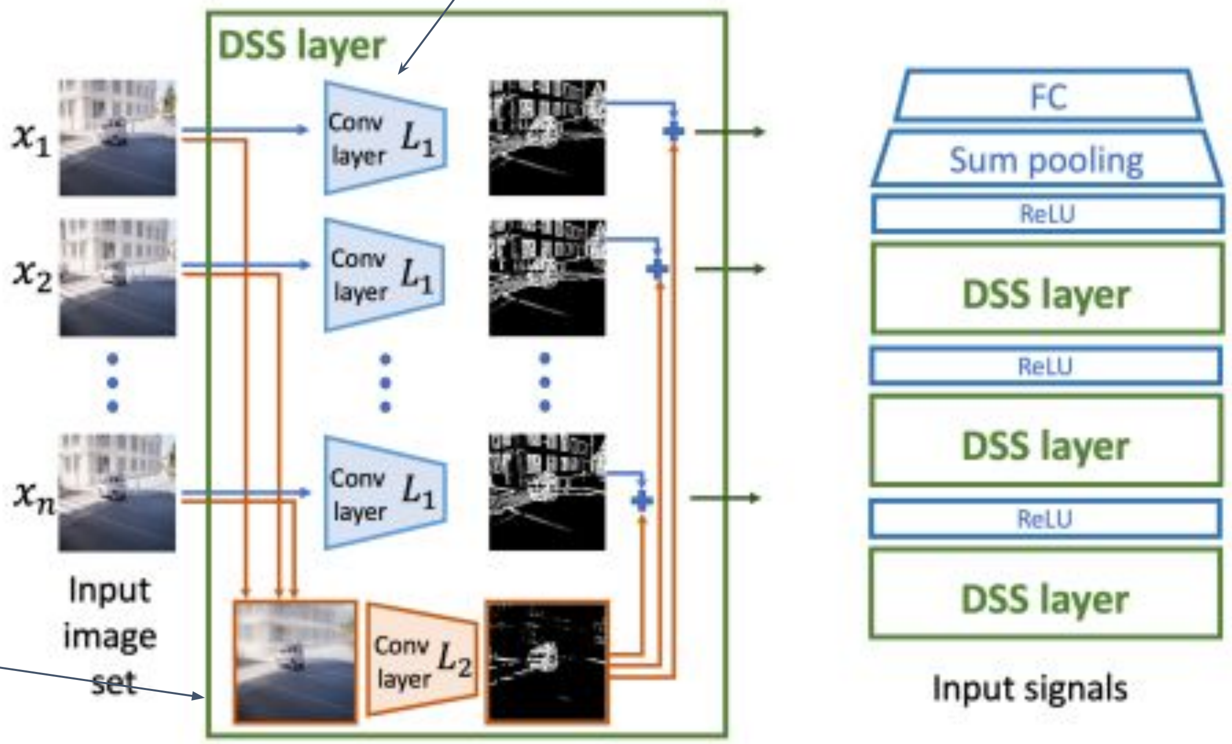
$G = H \times S_n$ - all possible permutations, i.e. translation invariance

S_n - order of input samples

The model

Shared weights

Sum of images



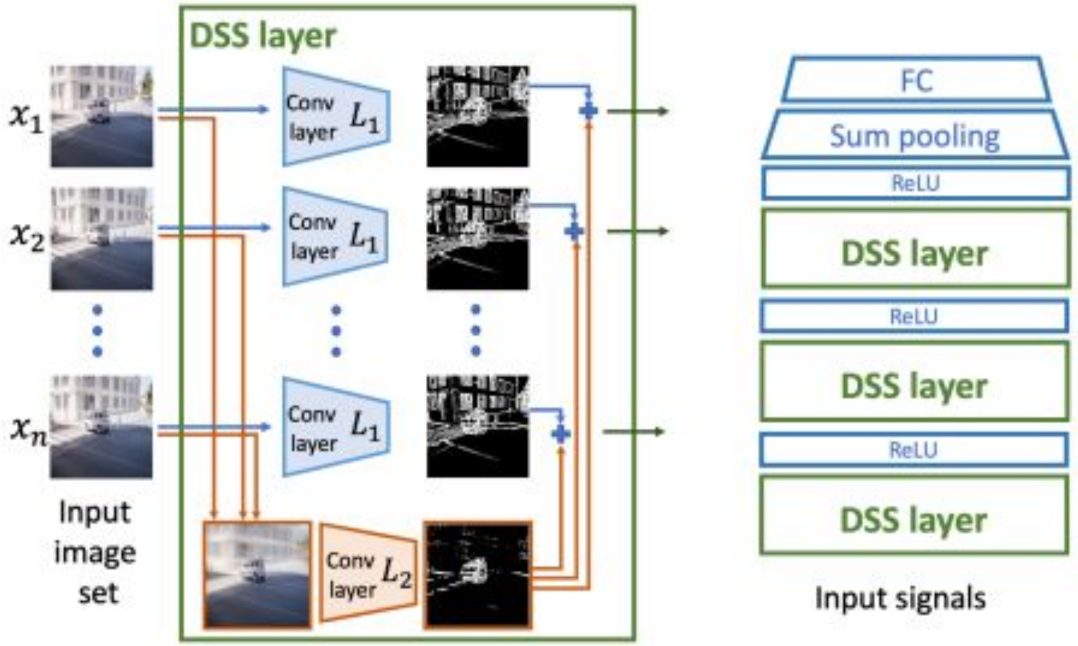
Deep Sets for Symmetric elements layers (DSS)

The model

Theorem 1. *Any linear $G-$equivariant layer $L : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$ is of the form*

$$L(X)_i = L_1^H(x_i) + L_2^H\left(\sum_{j \neq i}^{n} x_j\right),$$

where L_1^H, L_2^H are linear H-equivariant functions



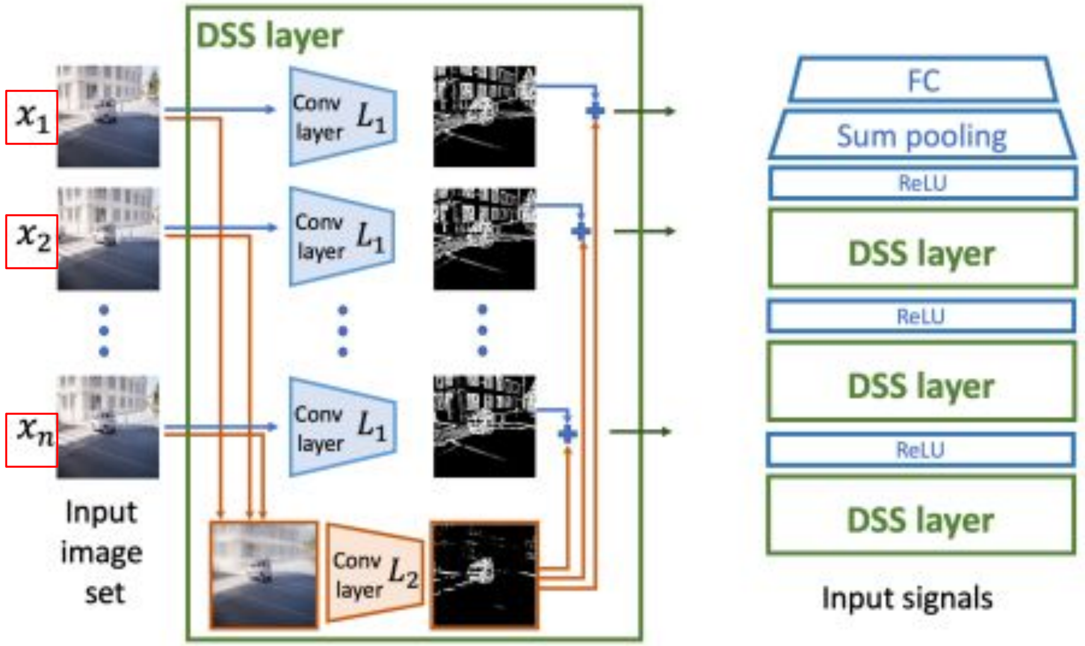
The model

Theorem 1. *Any linear $G-$equivariant layer $L : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$ is of the form*

DSS layer

$$L(X)_i = L_1^H(\boxed{x_i}) + L_2^H\left(\sum_{j \neq i}^{n} \boxed{x_j}\right),$$

where L_1^H, L_2^H are linear H-equivariant functions



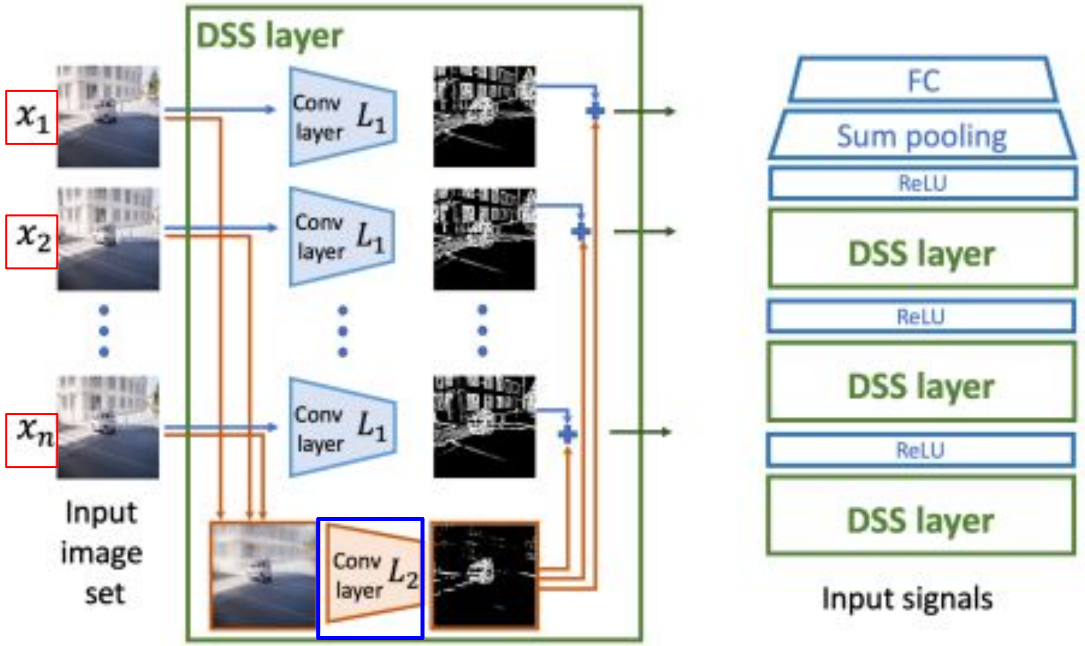
The model

Theorem 1. *Any linear $G-$equivariant layer $L : \mathbb{R}^{n\times d} \rightarrow \mathbb{R}^{n\times d}$ is of the form*

DSS layer

$$L(X)_i = L_1^H(\boxed{x_i}) + \boxed{L_2^H}\left(\sum_{j\neq i}^{n} \boxed{x_j}\right),$$

where L_1^H, L_2^H are linear H-equivariant functions



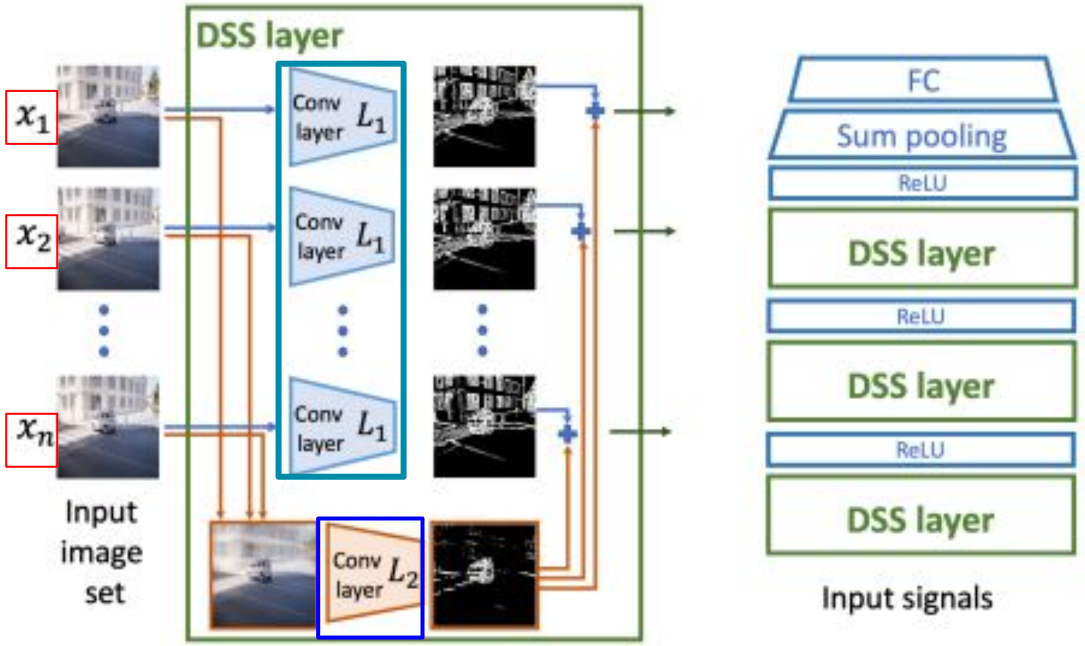
The model

Theorem 1. *Any linear G–equivariant layer $L : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ is of the form*

DSS layer

$$L(X)_i = \boxed{L_1^H}(\boxed{x_i}) + \boxed{L_2^H}\left(\sum_{j\neq i}^{n} \boxed{x_j}\right),$$

where L_1^H, L_2^H are linear H-equivariant functions



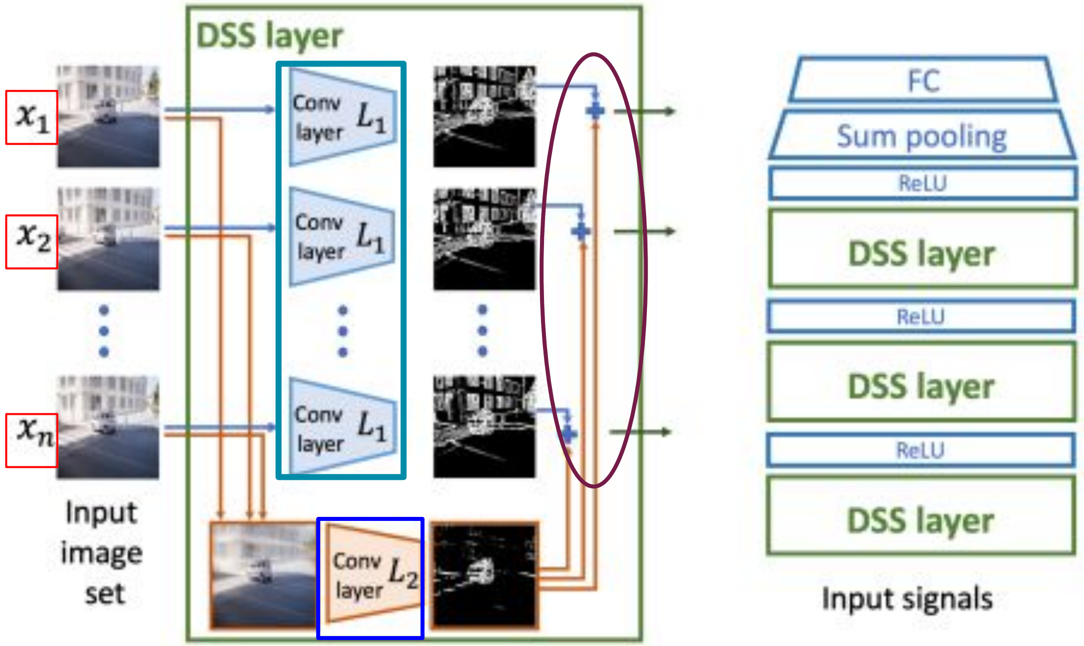
The model

Theorem 1. *Any linear $G-$equivariant layer $L : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ is of the form*

DSS layer

$$L(X)_i = L_1^H(x_i) + L_2^H\left(\sum_{j\neq i}^{n} x_j\right),$$

where L_1^H, L_2^H are linear H-equivariant functions

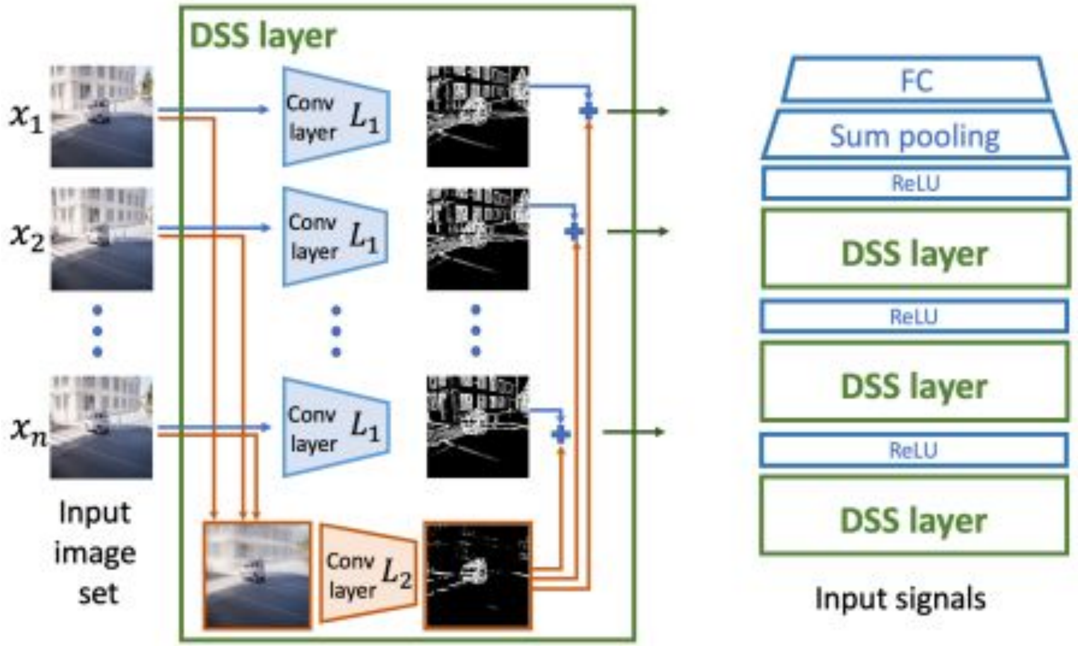


The model

Theorem 1. *Any linear G–equivariant layer $L : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$ is of the form*

$$L(X)_i = L_1^H(x_i) + L_2^H\left(\sum_{j \neq i}^{n} x_j\right),$$

where L_1^H, L_2^H are linear H-equivariant functions

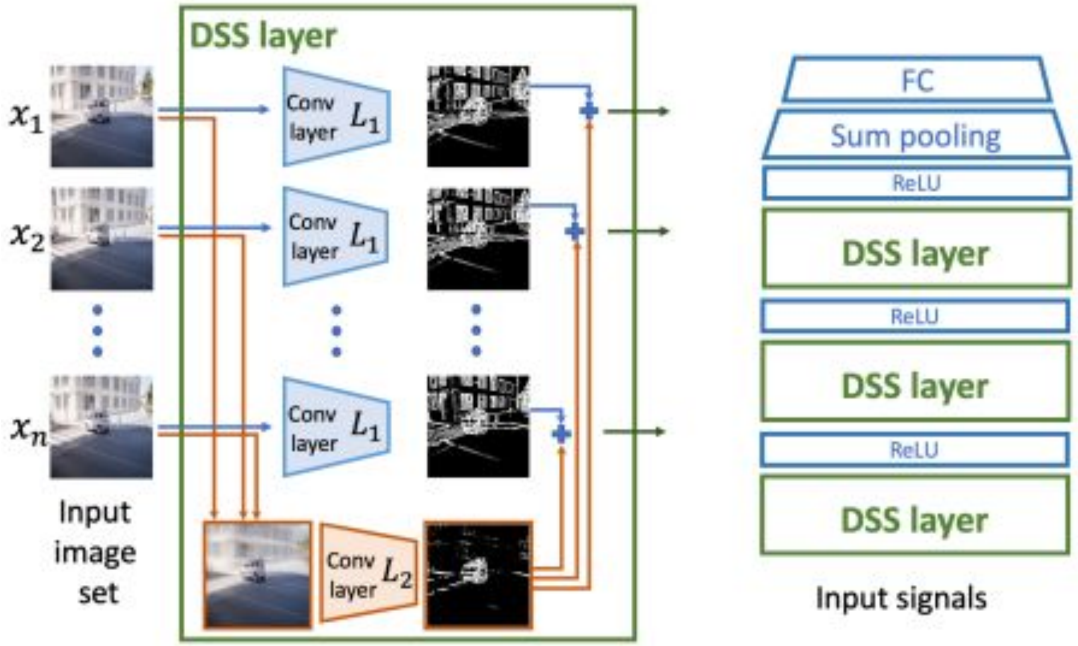


General intuition:

1. Sum of linear equivatiant functions are linear equivariant
2. Extension of rank of weight: E[G] = 2E[H] (improving the explicit power)

The model

Theorem 2. *Let $K \subset \mathbb{R}^{n\times d}$ be a compact domain such that $K = \cup_{g\in G} gK$ and $K \cap \mathcal{E} = \emptyset$. G-invariant networks are universal approximators (in $\|\cdot\|_\infty$ sense) of continuous G-invariant functions on K if H-invariant networks are universal*[1].

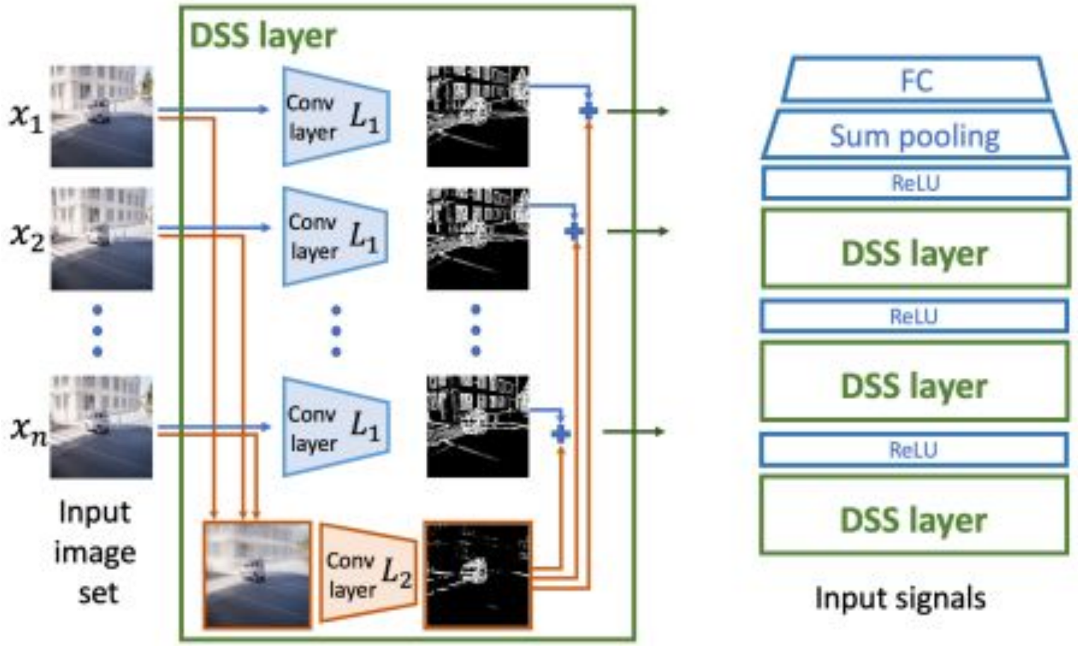


General intuition:

Sharing the information between L1 inputs to “synchronise independent H-equivariant outputs into the global G-H-equivariant representation”

The model

Theorem 2. *Let $K \subset \mathbb{R}^{n\times d}$ be a compact domain such that $K = \cup_{g\in G} gK$ and $K \cap \mathcal{E} = \emptyset$. G-invariant networks are universal approximators (in $\|\cdot\|_\infty$ sense) of continuous G-invariant functions on K if H-invariant networks are universal*[1].



General intuition:

Incorporating **global features** (extracted from **all pictures**) to the ***local features*** extracted from the ***each image***.

Experimental Results: overview

- Deep Sets (DS) : baseline, late aggregation
- DSS(sum): A sum aggregation, that corresponds to the Theorem 1
- DSS(max): A max aggregation
- DSS(Aittala): $L(x)_i \longrightarrow L^H(x_i) - (1/n)\sum^n_{j=1} L^H(x_j)$
- DSS(Sridhar): $L(x)_i \longrightarrow [L^H(x_i), \max^n_{j=1} L^H(x_j)]$, [] - concatenation

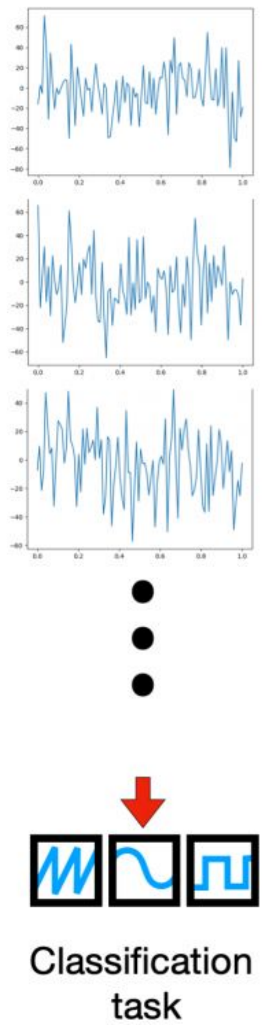
Experimental Results: 1D signals classification

- Problem: classification of the signal measured 25 times

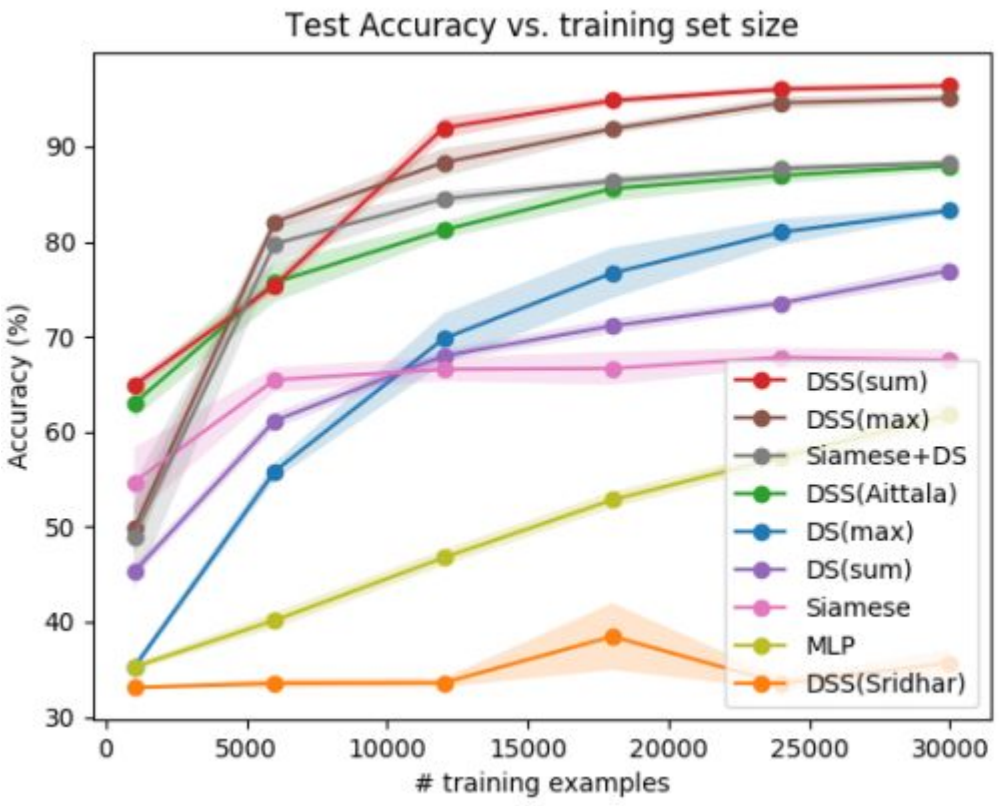
- Three types of waveforms: sine, triangle and square

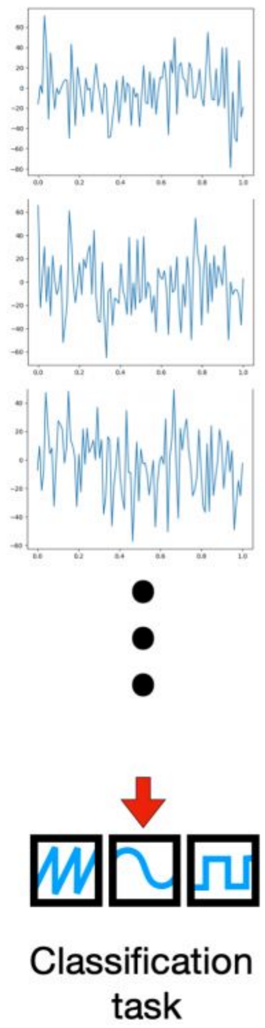
- For generating sets, random amplitude, DC shift, frequency, phase, noise were applied

- Base kernel: 1D convolution



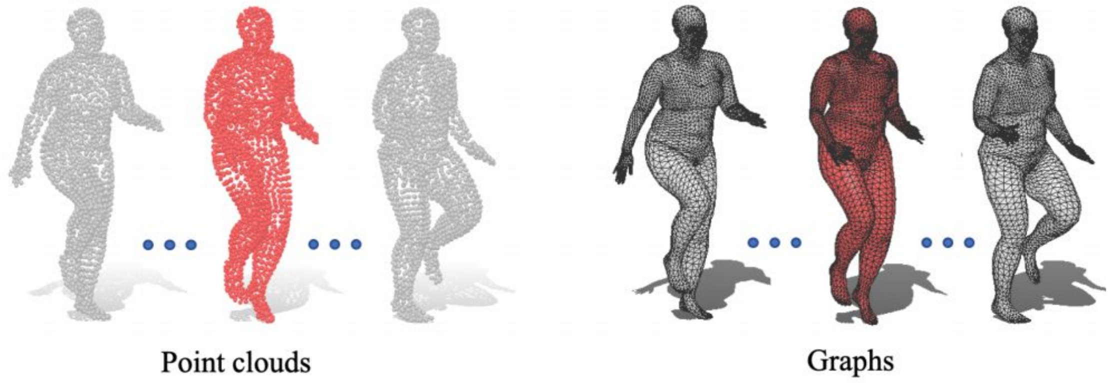
Experimental Results: 1D signals classification





Experimental Results: Frame selection from images and shapes

- Problem: selection the chronologically first/middle frame from the unordered sequence
- Data types: graphs (meshes), point clouds and images (video)



What’s the middle frame?

What’s the first frame?

- Base kernels: PointNet (pointcloud), GCN (meshes), 2D CNN (video/images)

Experimental Results: Frame selection from images and shapes

Dataset	Data type	Late Aggregation Siamese+DS	Early Aggregation DSS (sum)	DSS (max)	DSS (Sridhar)	DSS (Aittala)	Random choice
UCF101	Images	$36.41\% \pm 1.43$	$76.6\% \pm 1.51$	$76.39\% \pm 1.01$	$60.15\% \pm 0.76$	**77.96%** ± 1.69	12.5%
Dynamic Faust	Point-clouds	$22.26\% \pm 0.64$	$42.45\% \pm 1.32$	$28.71\% \pm 0.64$	**54.26%** ± 1.66	$26.43\% \pm 3.92$	14.28%
Dynamic Faust	Graphs	$26.53\% \pm 1.99$	$44.24\% \pm 1.28$	$30.54\% \pm 1.27$	**53.16%** ± 1.47	$26.66\% \pm 4.25$	14.28%

Experimental Results: Highest quality image selection

- Problem: selection the image with the best quality from the set of 20 images

- Images are generated by adding occlusion and Gaussian blur

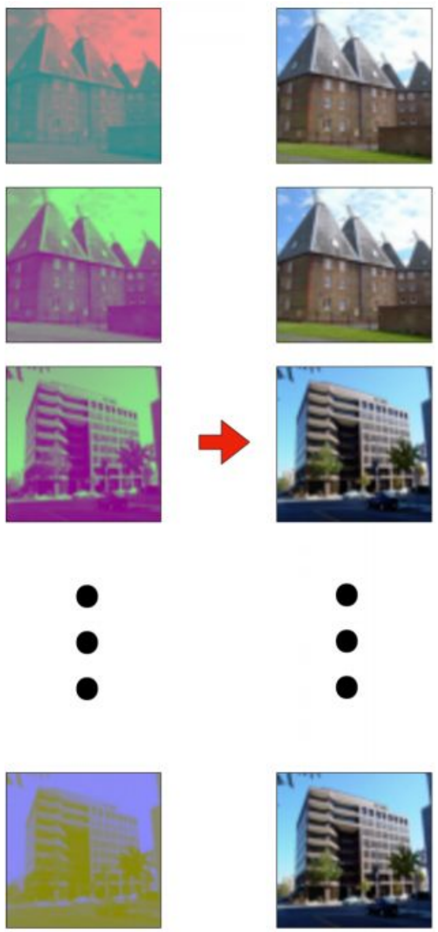
- Base kernel: 2D CNN

Experimental Results: Highest quality image selection

Noise type and strength	Late Aggregation	Early Aggregation				Random choice
	Siamese+DS	DSS (sum)	DSS (max)	DSS (Sridahr)	DSS (Aittala)	
Gaussian $\sigma = 10$	$77.2\% \pm 0.37$	**78.48%** ± 0.48	$77.99\% \pm 1.1$	$76.8\% \pm 0.25$	$78.34\% \pm 0.49$	5%
Gaussian $\sigma = 30$	$65.89\% \pm 0.66$	**68.35%** ± 0.55	$67.85\% \pm 0.40$	$61.52\% \pm 0.54$	$66.89\% \pm 0.58$	5%
Gaussian $\sigma = 50$	$59.24\% \pm 0.51$	**62.6%** ± 0.45	$61.59\% \pm 1.00$	$55.25\% \pm 0.40$	$62.02\% \pm 1.03$	5%
Occlusion 10%	$82.15\% \pm 0.45$	$83.13\% \pm 1.00$	**83.27** ± 0.51	$83.21\% \pm 0.338$	$83.19\% \pm 0.67$	5%
Occlusion 30%	$77.47\% \pm 0.37$	$78\% \pm 0.89$	$78.69\% \pm 0.32$	**78.71%** ± 0.26	$78.27\% \pm 0.67$	5%
Occlusion 50%	$76.2\% \pm 0.82$	**77.29%** ± 0.40	$76.64\% \pm 0.45$	$77.04\% \pm 0.75$	$77.03\% \pm 0.58$	5%

Experimental Results: Color channel matching

- Problem: combining randomly permuted color channels into the image
- Images generated from the Places & CelebA datasets
- Base kernel: 2D CNN


General Equivariant task

Experimental Results: Burst image deblurring

- Problem: combining 5 blurred images into the clean one

- Base kernel: 2D CNN

Experimental Results: Burst image deblurring

Trivial grayscale prediction

Task	Late Aggregation Siamese+DS	Early Aggregation DSS (sum)	DSS (max)	DSS (Sridahr)	DSS (Aittala)	TP
Color matching (places)	8.06 ± 0.06	1.78 ± 0.03	1.92 ± 0.07	1.97 ± 0.02	$\mathbf{1.67} \pm 0.06$	14.68
Color matching (CelebA)	6 ± 0.13	1.27 ± 0.07	1.34 ± 0.07	1.35 ± 0.03	$\mathbf{1.17} \pm 0.04$	18.72
Burst deblurring (Imagenet)	6.15 ± 0.05	6.11 ± 0.08	5.87 ± 0.05	21.01 ± 0.08	$\mathbf{5.7} \pm 0.13$	16.75

Median over all pixels in the set of images

Discussion of results

- DSS(max/sum) works all on all tasks, which makes it more universal compared to the DSS(Sridahr) and DSS(Aittala) implementations
- Early aggregation methods significantly overperformed Deep sets on point clouds and mesh data types, which could be related to incorporation of the global features into local representations
- DSS (Sridahr) works DSS (Aittala) aggregation techniques may fail

Critique

1. Theorem 1: assumption of the linearity of the H-equivariant functions
2. Based on #1, G-Equivariance is relying on the same position of objects in all images for deblurring tasks, if objects are shifted -> could be even worse
3. No consistency in experiments: different aggregation methods shown different results, sometimes even worse compared to the baseline

Contributions (Recap)

- Formulated the problem of limited expressive power of Deep Sets architectures
- Proposed a model architecture called DSS, both invariant and equivariant to the order/translation of input samples
- Mathematically proved of the proposed model architecture
- Provided a benchmarks on a variety sets of tasks and various data types

Thank you!