Reading:
Equivariant Multi- View Networks

Carlos Esteves* , Yinshuang Xu* , Christine Allen- Blanchette, Kostas Daniilidis (UPenn)

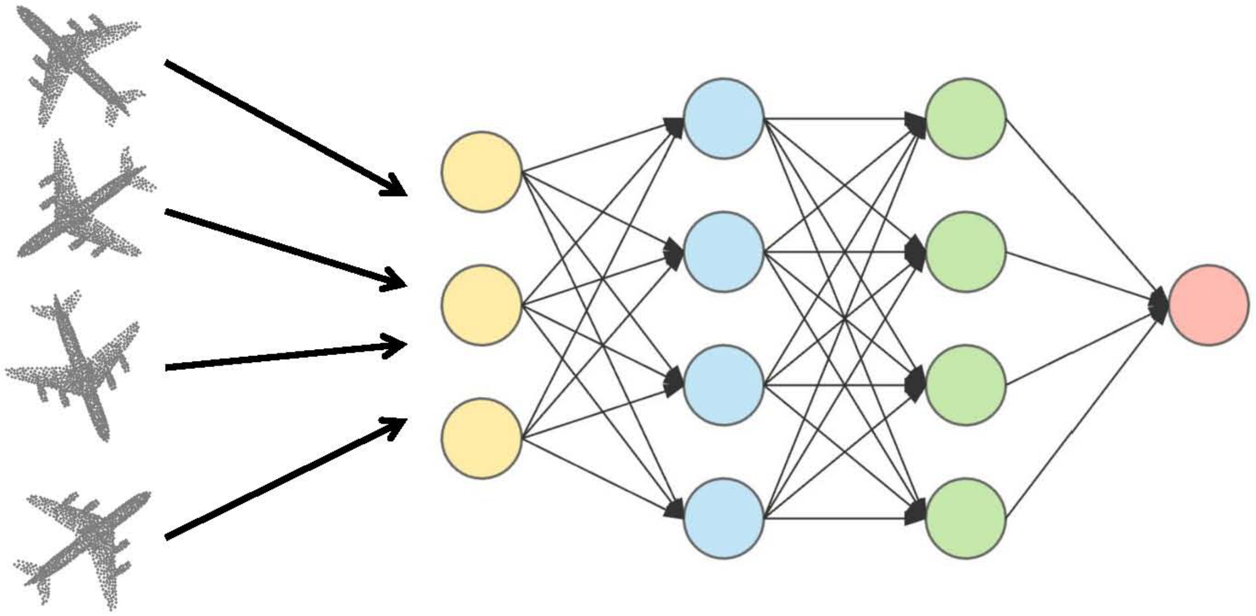
Presented by: Congyue Deng

Intuitions

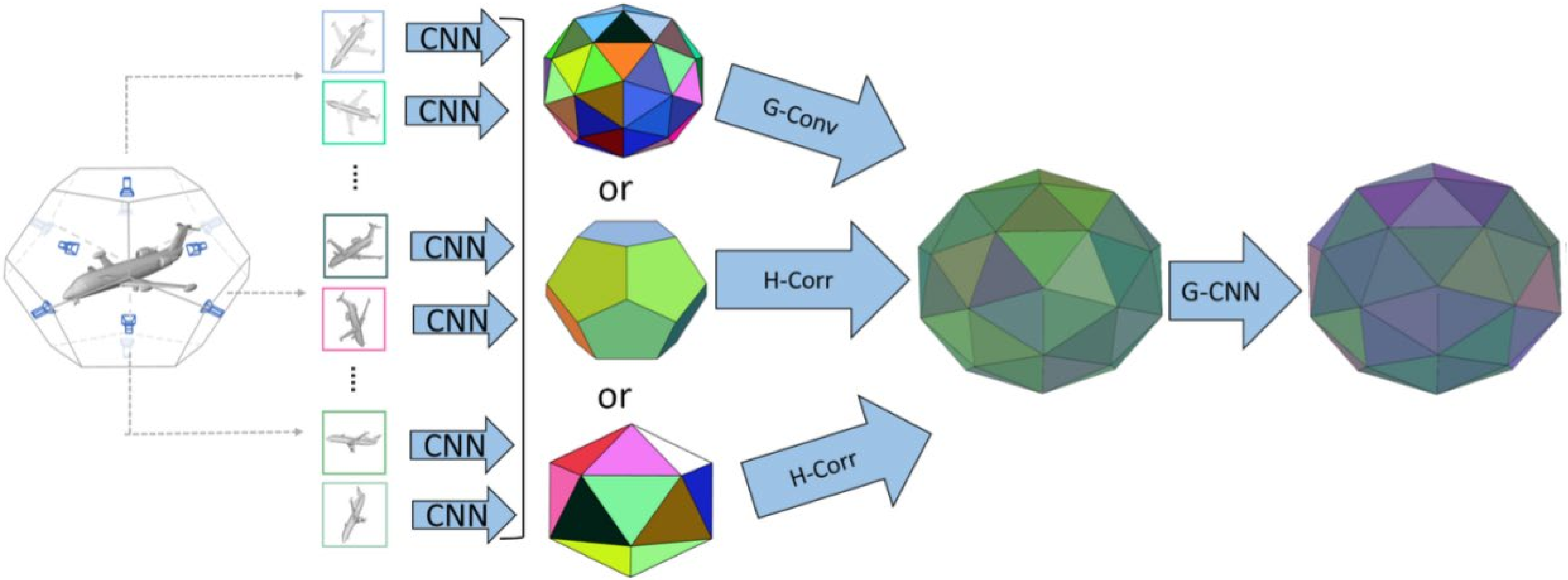
Feed multiple poses of the same object to the network **at once**

"If we don't know what pose to look at, why not just look at **all** poses!"

- Can reproduce fairly good results
- Sacrifice data-efficiency – more memory consumption
- Theoretically equivariant – up to precision errors caused by discretization



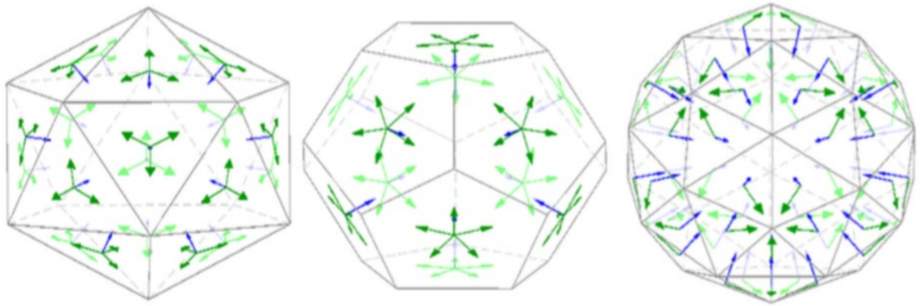
Idea overview



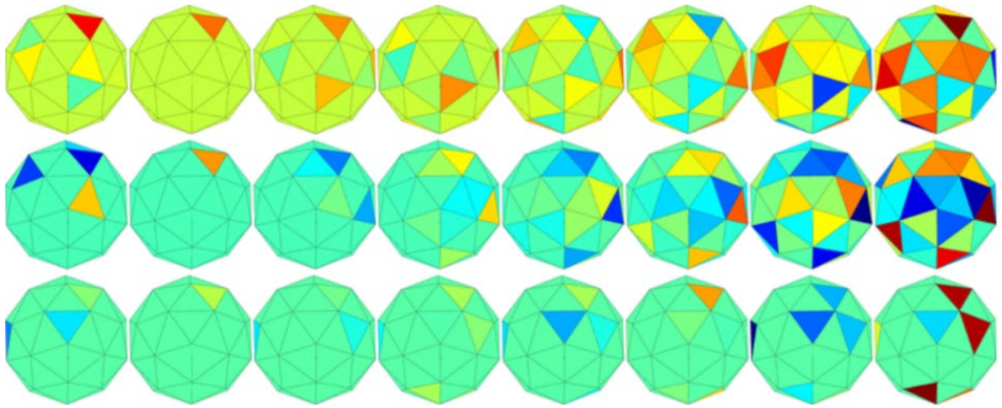
Let's build the network!

We need to…

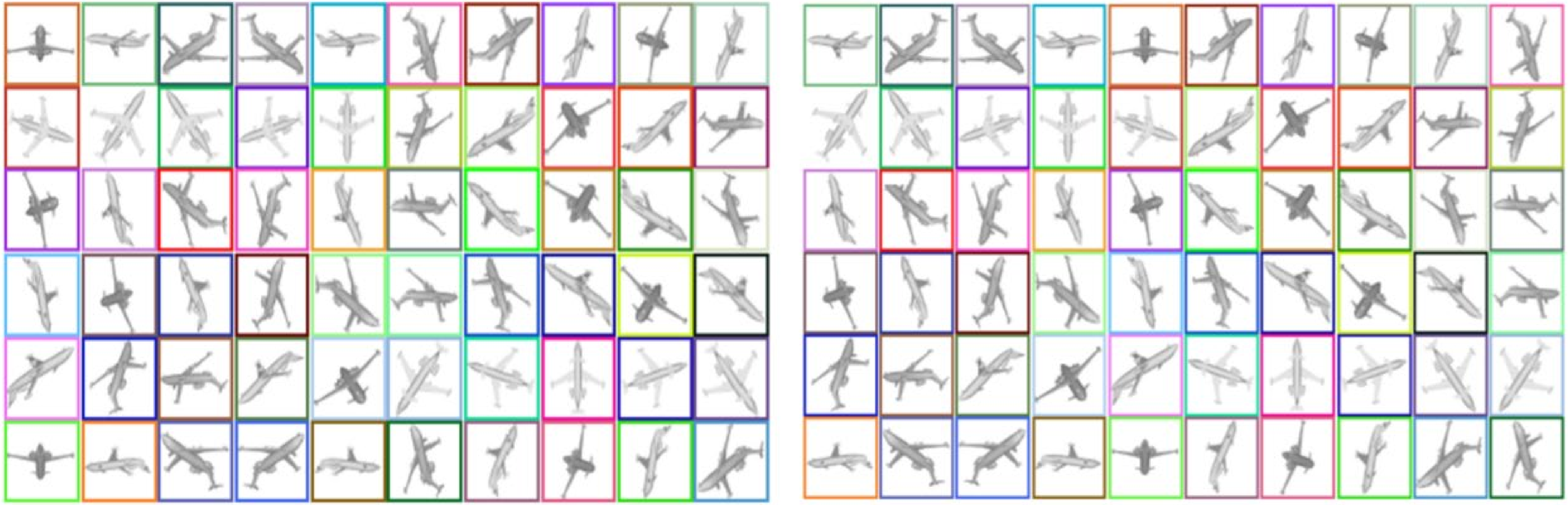
- **Step 1:** Uniformly sample camera views



- **Step 2:** Define group convolution filters



Step 1: Sample views



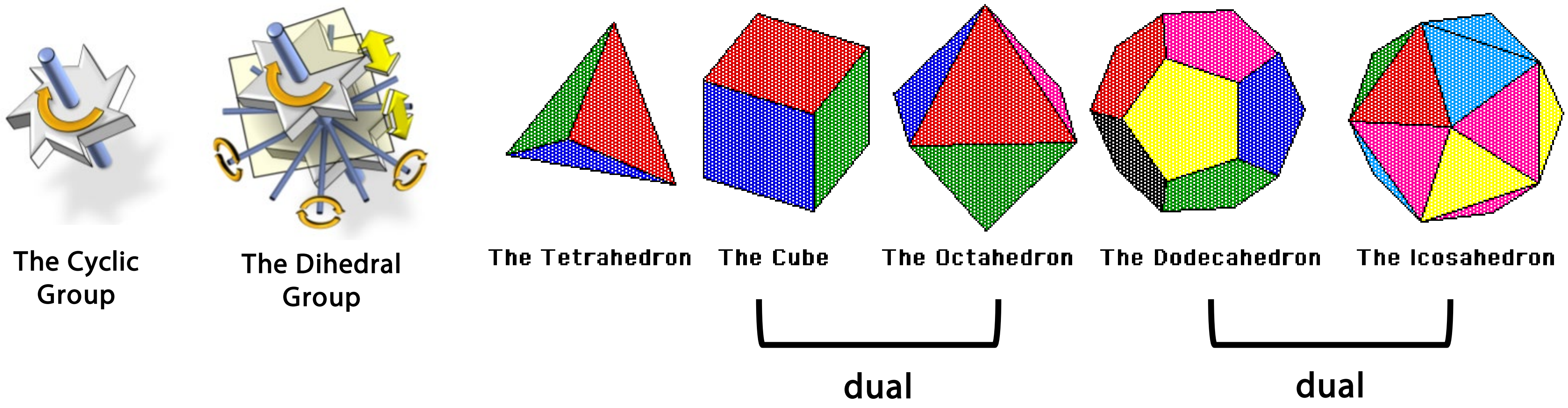
Step 1: Sample views

Can also sample **panoramic images!**

Finite 3D rotation groups

These are the only finite subgroups of SO(3)

- Equivalently, SU(2)

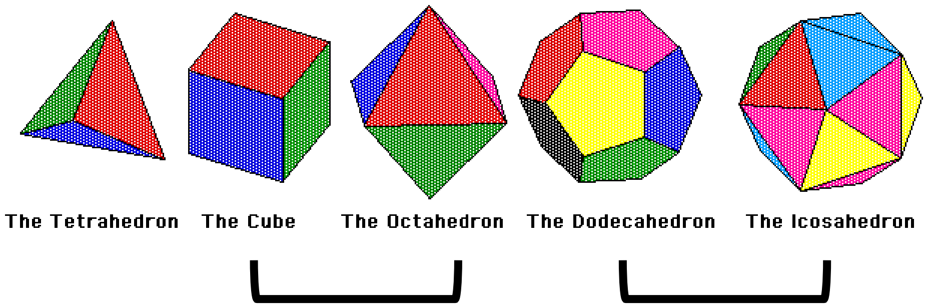


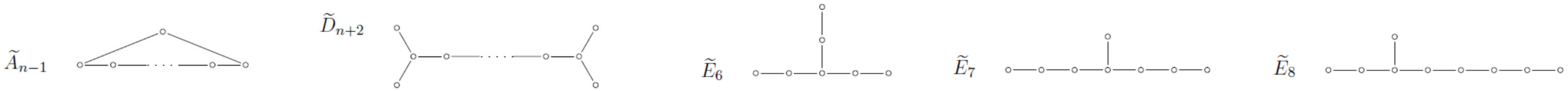
Finite 3D rotation groups — fun fact

They correspond to a lot of fancy things in algebra (representation theory, category theory, ..) — the McKay Correspondence

The Cyclic Group

The Dihedral Group





Even Wolf Prize winner studies this

Affine quivers and canonical bases

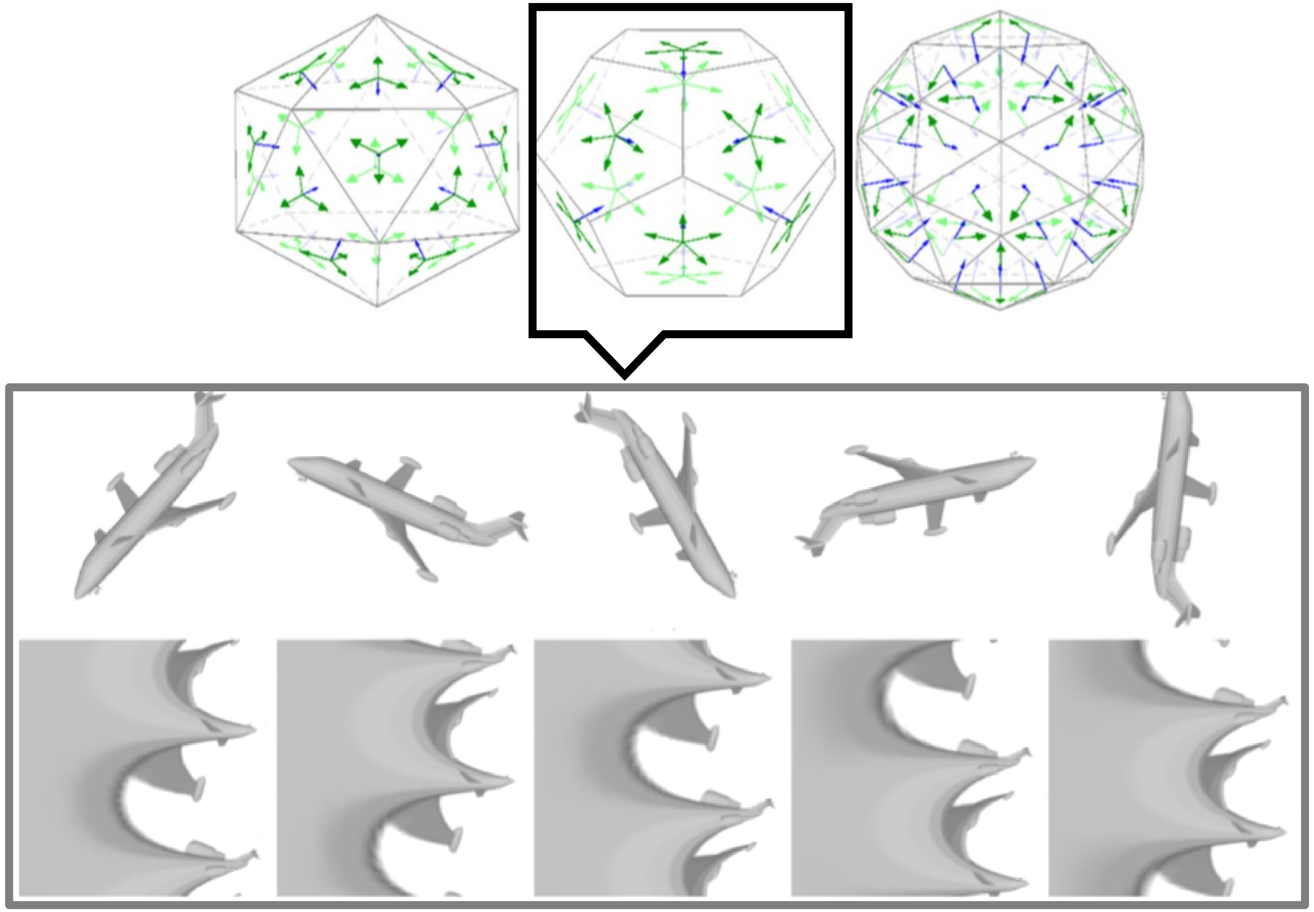
Lusztig, George

Publications Mathématiques de l'IHÉS, Tome 76 (1992), pp. 111-163.

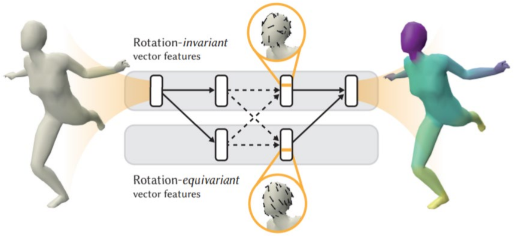
Equivariance with fewer views

In-plane rotations

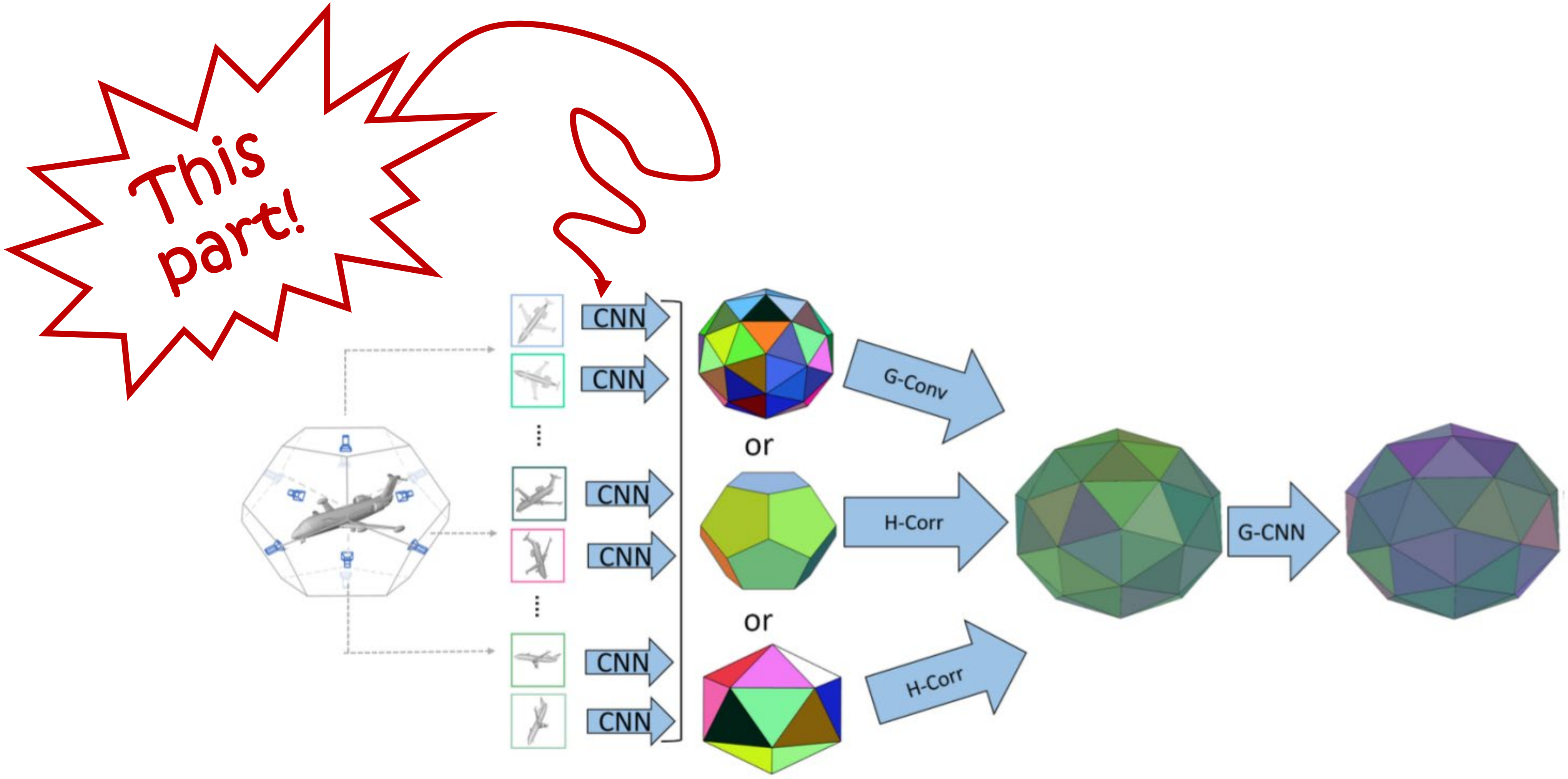
- can be cancelled out by the **equivariance of 2D convolution** on images



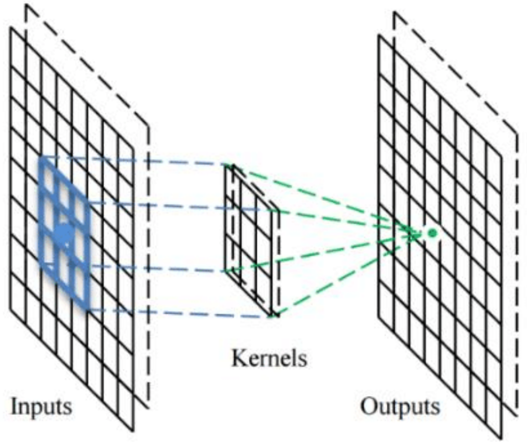
Is somehow a discrete version of
https://arxiv.org/pdf/2006.01570.pdf



Equivariance with fewer views



Group convolution



Conventional 2D convolution

$$(f * h)(y) = \int_{x \in \mathbb{R}^2} f(x)h(y-x)\,dx$$

- Equivariant to the 2D translation group

Generalize to **arbitrary groups**

$$(f * h)(y) = \int_{g \in G} f(g)h(g^{-1}y)\,dg$$

Discrete version

$$f_j^{\ell+1}(y) = \sigma\left(\sum_{i=1}^{c_i}\sum_{g\in G} f_i^{\ell}(g)h_{ij}(g^{-1}y)\right)$$

Group convolution

Convolution on **homogeneous spaces**

$$(f * h)(y) = \int_{g \in G} f(g\eta) h(g^{-1}y)\, dg$$ ← homogeneous space convolution **(H-Conv)**

$$(f \star h)(g) = \int_{x \in \mathcal{X}} f(gx) h(x)\, dx$$ ← homogeneous space correlation **(H-Corr)**

Intuitively, a homogeneous space is a group without a specified "0"

- e.g. a 2D plane on which you don't know where the origin is

Compare with convolution on groups $(f * h)(y) = \int_{g \in G} f(g) h(g^{-1}y)\, dg$

Group convolution

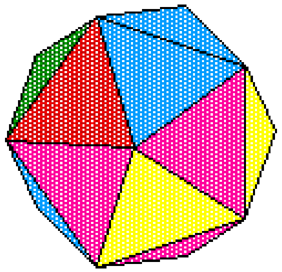
Convolution on **homogeneous spaces**

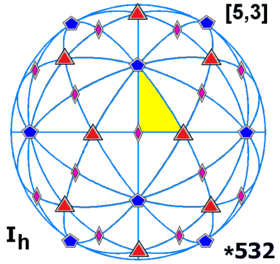
- Discrete version

$$ {}^{*}f_j^{\ell+1}(y) = \sigma\left(\sum_{i=1}^{c_i}\sum_{g\in G} f_i^{\ell}(g\eta)h_{ij}(g^{-1}y)\right) $$

$$ {}^{\star}f_j^{\ell+1}(g) = \sigma\left(\sum_{i=1}^{c_i}\sum_{x\in \mathcal{X}} f_i^{\ell}(gx)h_{ij}(x)\right) $$

Why homogeneous space? — So we can operate on the primitives instead of the abstract groups

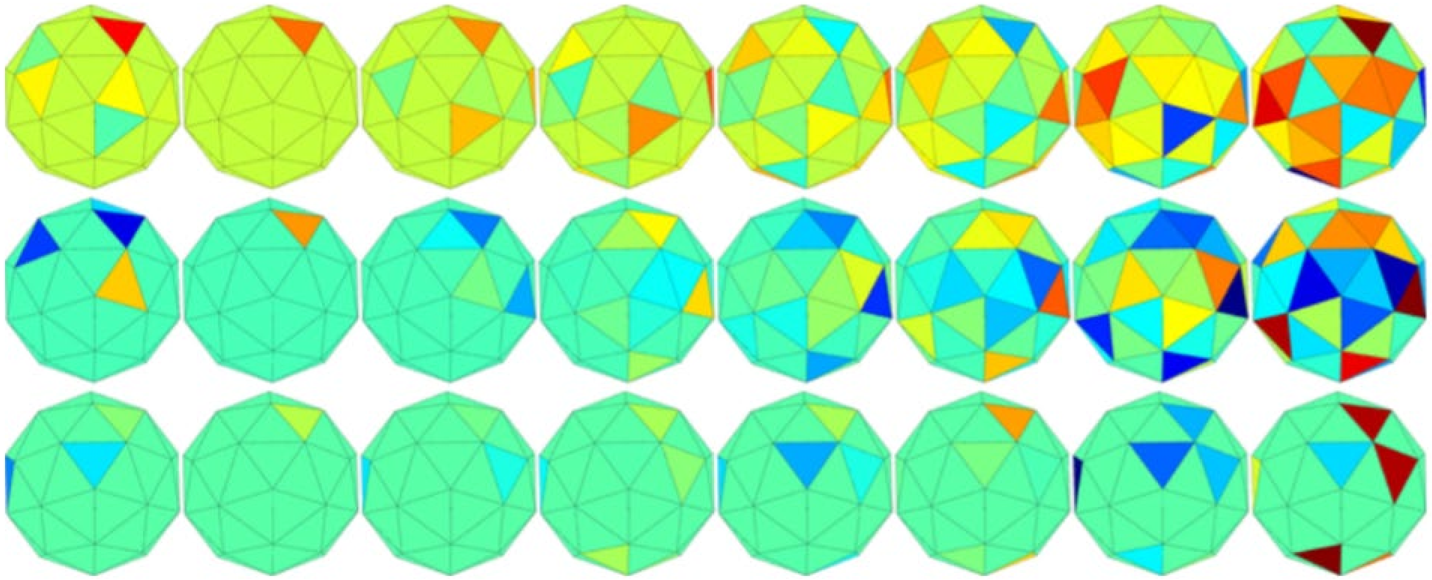




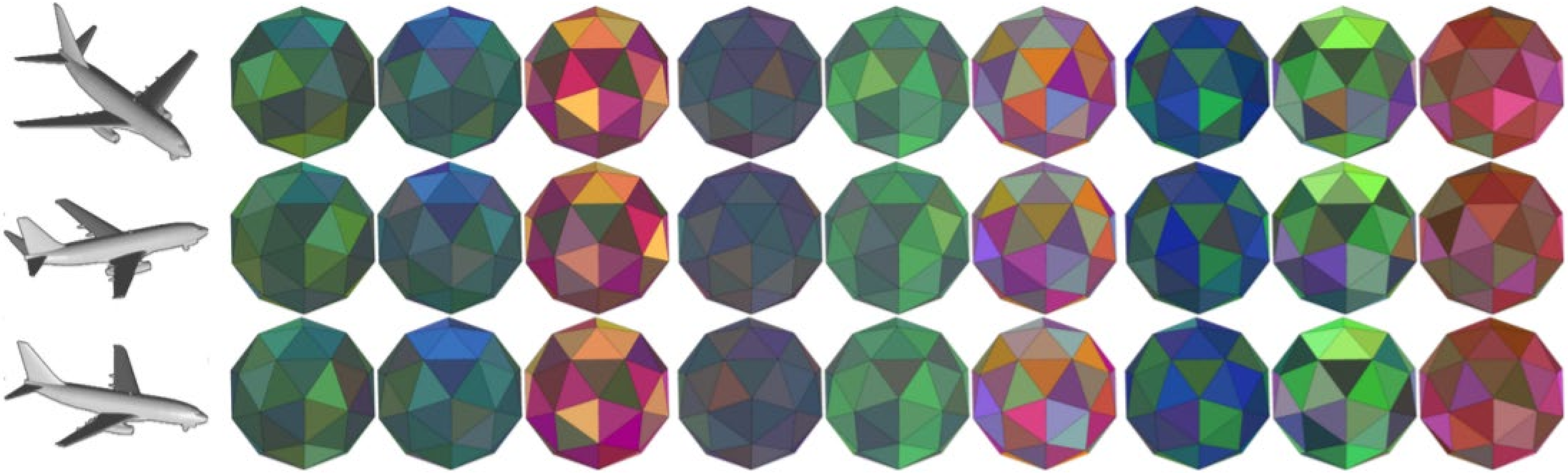
Filter localization

Choose a subset S of G that is allowed to have nonzero filter values while $h(G - S)$ is set to zero

- S is a fixed hyperparameter



Feature visualizations



Experimental results

ModelNet40 classification & retrieval

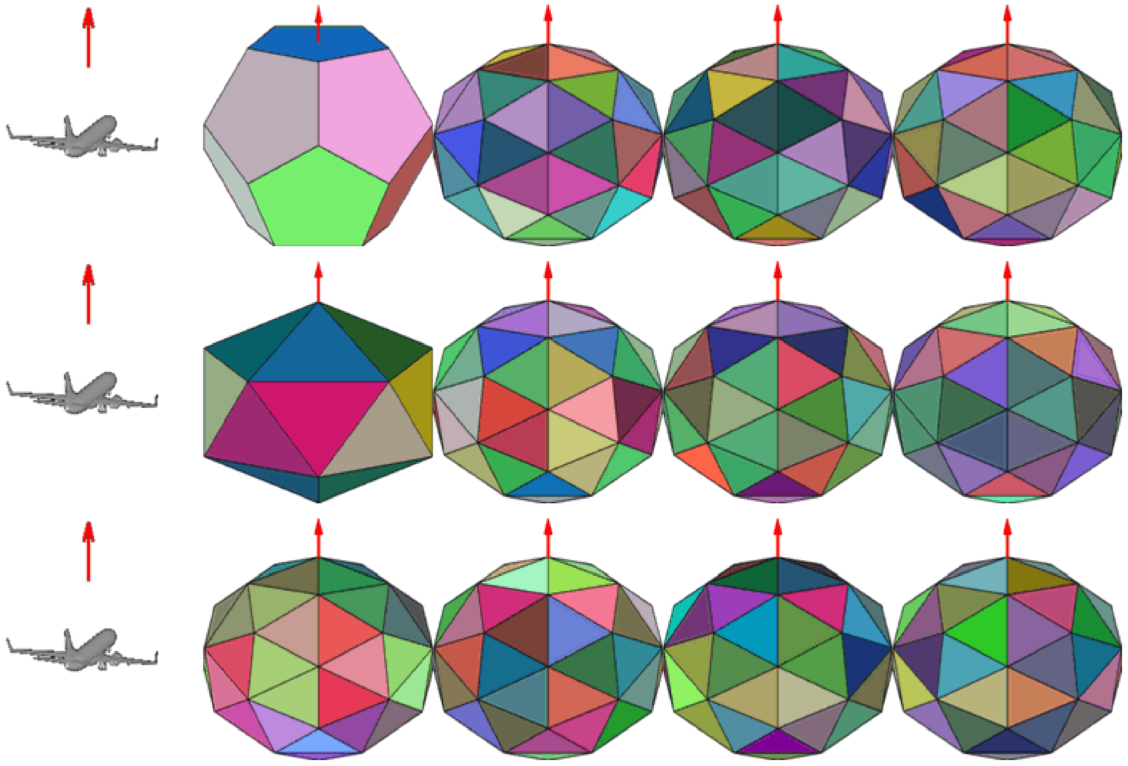
	M40 (aligned)		M10 (aligned)	
	acc	mAP	acc	mAP
MVCNN-12 [33]	90.1	79.5	-	-
SPNet [40]	92.63	85.21	*97.25*	94.20
PVNet [41]	93.2	89.5	-	-
SV2SL [14]	93.40	89.09	94.82	91.43
PANO-ENN [31]	**95.56**	86.34	96.85	93.2
MVCNN-M-12	94.47	89.13	96.33	93.54
Ours-12	94.51	*91.82*	96.33	95.30
Ours-20	*94.69*	91.42	**97.46**	*95.74*
Ours-60	94.36	91.04	96.80	95.25
Ours-R-12	94.67	**93.56**	96.78	**96.18**

	M40 (rotated)	
	acc	mAP
MVCNN-80 [33]	86.0	-
RotationNet [20]	80.0	74.20
Spherical CNN [6]	86.9	-
MVCNN-M-60	90.68	78.18
Ours-12	88.50	79.58
Ours-20	89.98	80.73
Ours-60	*91.00*	*82.61*
Ours-R-60	**91.08**	**88.57**

Experimental results

ModelNet40
classification
& retrieval

- More coherent across rotations
- But still not fully equivariant — a small performance gap from aligned to rotated shapes
- Sampling more views alleviates the issue



Thank you!