Supplementary material: Equivariant Multi-View Networks

We divide this material in "Extended experiments", "Proofs", "Implementation details", and "Visualization". Numbering and citations follow the main text.

A. Extended experiments

We include an ablation experiment and further results for ModelNet, SHREC'17 large scale retrieval, comparison with RotationNet [20], and Matterport3D scene classification.

A.1. Ablation

We run an experiment to compare effects of (1) filter support size, (2) number of G-Conv layers, and (3) missing views. We evaluate on rotated ModelNet40 with "Ours-60" model as baseline that has with 9 elements in the support, 3 G-Conv layers and all 60 views.

When considering less than 60 views, we introduce view dropout during training where a random number (between 1 and 30) of views is selected for every mini-batch. This improves robustness to missing views. During test, a fixed number of views is used. Table 5 shows the results. As expected, we can see some decline in performance with fewer layers and smaller support, which reduces the receptive field at the last layer. Our method is shown to be robust to missing up to 50% of the views, with noticeable drop in performance when missing 80% or more.

In the second ablation experiment, we investigate the performance as the assumption that the views can be associated with group elements is broken. We perturb the camera pose for each view with randomly sampled rotations given some standard deviation. The model is trained once on ModelNet40 (aligned) and tested under different levels of perturbations. Table 6 shows the results.

support	layers	views	pretrained	acc	mAP
9	3	60	yes	91.00	82.61
6	3	60	yes	90.63	81.90
3	3	60	yes	89.74	80.49
9	2	60	yes	91.00	81.47
9	1	60	yes	90.88	79.59
9	3	30	yes	89.50	79.20
9	3	10	yes	88.32	74.65
9	3	5	yes	82.77	64.88
9	3	60	no	87.40	70.44

Table 5: Ablation study on rotated ModelNet40. Our best performing model is on the top row.

std [deg]	acc	mAP
0	93.96	89.74
5	94.20	89.34
15	92.70	86.97
30	88.61	81.16
45	80.98	71.61

Table 6: We perturb the camera pose for each view to gradually break the assumption that they form a group. The model is trained with no perturbations and tested under different levels of perturbations.

A.2. SHREC'17

We show all metrics for the SHREC'17 large scale retrieval challenge in Table 7.

A.3. ModelNet

Since some methods show ModelNet40 results as averages per class instead of the more common average per instance, we include exteded tables with these metrics. We also present results on rotated ModelNet10. Tables 8 and 9 shows the results for the aligned and rotated versions, respectively.

A.4. Comparison with RotationNet

We provide further comparison against RotationNet [20]. While RotationNet remains SoTA on aligned ModelNet classification, our method is superior on all retrieval benchmarks. We also outperform RotationNet on more challenging classification taks: rotated and aligned ShapeNet, and rotated ModelNet. Table 10 shows the results.

A.5. Scene classification

We show examples of original input and our 12 overlapping views in Figure 6. The complete set of results in the same format as [2] is shown in Table 11.

B. Proofs

We demonstrate the equivariance of operations, and show that MVCNN [33] is a special case of our method.

B.1. Equivariance of G-Conv, H-Conv, H-Corr

We demonstrate here the equivariance of the main operations used in our method: group convolution, homogeneous space convolution and correlation. We assume a compact group G and one of its homogeneous spaces $\mathcal{X}$. $\mathcal{T}_k$ is the action of $k \in G$. Let us start with G-Conv (3), where

		micro					macro				
Method	score	P@N	R@N	F1@N	mAP	G@N	P@N	R@N	F1@N	mAP	G@N
RotatNet [20]	67.8	81.0	80.1	79.8	77.2	86.5	60.2	63.9	59.0	58.3	65.6
ReVGG [29]	61.8	76.5	80.3	77.2	74.0	82.8	51.8	60.1	51.9	49.6	55.9
DLAN [13]	57.0	81.8	68.9	71.2	66.3	76.2	61.8	53.3	50.5	47.7	56.3
MVCNN-12 [33]	65.1	77.0	77.0	76.4	73.5	81.5	57.1	62.5	57.5	56.6	64.0
MVCNN-M-12	69.1	83.1	77.9	79.4	74.9	83.8	*66.8*	68.4	65.2	63.2	70.3
Ours-12	70.7	83.1	80.5	81.1	77.7	86.3	65.3	68.7	64.8	63.6	70.8
Ours-20	71.4	*83.6*	*80.8*	*81.5*	*77.9*	*86.8*	66.4	*70.1*	65.9	64.9	71.9
Ours-60	*71.7*	**84.0**	80.5	81.4	77.8	86.4	**67.1**	**70.7**	**66.6**	**65.6**	**72.3**
Ours-R-20	**72.2**	*83.6*	**81.7**	**82.0**	**79.1**	**87.5**	*66.8*	69.9	*66.1*	*65.4*	**72.3**
DLAN [13]	56.6	**81.4**	68.3	70.6	65.6	75.4	**60.7**	53.9	50.3	47.6	56.0
ReVGG [29]	55.7	70.5	**76.9**	71.9	*69.6*	78.3	42.4	56.3	43.4	41.8	47.9
RotatNet [20]	46.6	65.5	65.2	63.6	60.6	70.2	37.2	39.3	33.3	32.7	40.7
MVCNN-80 [33]	45.1	63.2	61.3	61.2	53.5	65.3	40.5	48.4	41.6	36.7	45.9
MVCNN-M-60	57.5	77.7	67.6	71.1	64.1	75.9	55.7	56.9	53.5	50.9	59.7
Ours-12	58.1	76.1	70.0	72.0	66.4	76.7	54.6	55.7	52.6	49.8	58.6
Ours-20	59.3	76.4	70.5	72.4	66.9	77.0	54.6	58.0	53.7	51.7	60.2
Ours-60	*62.1*	*78.7*	72.9	*74.7*	*69.6*	*79.6*	57.6	*60.1*	*56.3*	*54.6*	*63.0*
Ours-R-60	**63.5**	*78.7*	*75.0*	**75.9**	**71.8**	**81.1**	*58.3*	**60.6**	**56.9**	**55.1**	**63.3**

Table 7: SHREC'17 retrieval results. Top block: aligned dataset; bottom: rotated. Methods are ranked by the micro and macro mAP average (namely, the "score" in the second column). We also show Precision (P), Recall (R), F-score (F1), mean average precision (mAP) and normalized discounted cumulative gain (G), where N is the number of retrieved elements.

	M40 (aligned)				M10 (aligned)			
	acc inst	acc cls	mAP inst	mAP cls	acc inst	acc cls	mAP inst	mAP cls
Ours-12	94.51	92.49	91.82	88.28	96.33	96.00	95.30	95.00
Ours-20	94.69	92.56	91.42	87.71	97.46	97.34	95.74	95.58
Ours-60	94.36	92.40	91.04	87.30	96.80	96.58	95.25	95.01
Ours-R-20	94.44	92.49	93.19	89.65	97.02	96.97	96.59	96.46

Table 8: Aligned ModelNet results. We include classification accuracy and retrieval mAP per class (cls) and per instance (inst).

$f, h\colon G \mapsto \mathbb{R}$:

$$\begin{aligned}(\mathcal{T}_k f * h)(y) &= \int_{g\in G} f(k^{-1}g)h(g^{-1}y)\,dg \\ &= \int_{g\in G} f(l)h((kl)^{-1}y)\,dl \\ &= \int_{g\in G} f(l)h((l^{-1}k^{-1}y)\,dl \\ &= (f*h)(k^{-1}y) \\ &= \mathcal{T}_k(f*h)(y).\end{aligned}$$

For H-Conv (4), where $f, h\colon \mathcal{X} \mapsto \mathbb{R}$, we have:

$$\begin{aligned}(\mathcal{T}_k f * h)(y) &= \int_{g\in G} f(k^{-1}g\eta)h(g^{-1}y)\,dg \\ &= \int_{g\in G} f(l\eta)h((kl)^{-1}y)\,dl \\ &= \int_{g\in G} f(l\eta)h(l^{-1}k^{-1}y)\,dl \\ &= (f*h)(k^{-1}y) \\ &= \mathcal{T}_k(f*h)(y).\end{aligned}$$

	M40 (rotated)				M10 (rotated)			
	acc inst	acc cls	mAP inst cls	mAP cls	acc inst	acc cls	mAP inst	mAP cls
Ours-12	88.50	85.77	79.58	74.64	91.89	91.54	86.93	86.08
Ours-20	89.98	87.65	80.73	75.65	92.60	92.35	87.27	86.65
Ours-60	91.00	89.24	82.61	78.02	92.83	92.80	88.47	88.02
Ours-R-20	91.08	88.94	88.57	84.37	93.05	93.08	92.07	91.99

Table 9: Rotated ModelNet results. We include ModelNet10 and classification accuracy and retrieval mAP per class (cls) and per instance (inst).

Figure 6: Top: original input from MatterPort3D [2] scene classification task. Bottom: our set of 12 overlapping views.

	M40 (al)		S17 (al)		S17 (rot)	
	acc	mAP	acc	mAP	acc	mAP
RotNet	**97.37**	93.00	85.39	67.8	77.37	46.6
Ours	94.67	**93.56**	**89.15**	**72.2**	**85.93**	**63.5**

Table 10: Classification accuracy (acc) and retrieval (mAP) comparison against RotationNet. Results for aligned (al) and rotated (rot) datasets, and for the SHREC'17 split of ShapeNet (S17). The mAP for SHREC'17 is the average between micro and macro (score).

Finally, for H-Corr (5), where $f, h\colon \mathcal{X} \mapsto \mathbb{R}$, we have:

$$\begin{aligned}(\mathcal{T}_k f \star h)(g) &= \int_{x \in \mathcal{X}} f(k^{-1}gx)h(x)\,dx \\ &= (f \star h)(k^{-1}g) \\ &= \mathcal{T}_k^{'}(f \star h)(y).\end{aligned}$$

Note that, in this case, $\mathcal{T}_k^{'}$ is not equal $\mathcal{T}_k$ because inputs and outputs are in different spaces.

B.2. MVCNN is a special case

Now we show that our model can replicate MVCNN [33] by fixing the filters $h_{ij}(g)$, where i, j denote the output and input channel and g denotes the element in group:

$$h_{ij}(g) = \begin{cases} 1 & i = j \quad \text{and} \quad g = e \\ 0 & \text{else} \end{cases} \tag{10}$$

Combining with group correlation (the result can also be achieved by group convolution), we show that

	avg.	office	lounge	family room	entryway	dining room	living room	stairs	kitchen	porch	bathroom	bedroom	hallway
single [2]	33.3	20.3	**21.7**	16.7	1.8	20.4	27.6	49.5	52.1	57.4	44.0	43.7	44.7
pano [2]	41.0	*26.5*	15.4	11.4	3.1	27.7	34.0	60.6	55.6	62.7	65.4	62.9	66.6
MVCNN-M-12	*51.9*	18.0	*16.4*	*23.8*	*8.6*	*46.7*	*37.1*	**84.1**	*73.3*	**81.0**	**78.2**	*81.7*	**73.8**
Ours-12	**53.8**	**27.9**	*16.4*	**33.3**	**11.4**	**51.1**	**41.3**	*80.4*	**75.8**	*79.0*	*72.5*	**82.9**	*73.5*

Table 11: Matterport3D panoramic scene classification extended results.

$$(f \star h)_i(k) = \sum_{j=1}^{c_1} \sum_{g \in G} f_j(kg) h_{ij}(g) = \begin{cases} f_i(k) & 1 \leq i \leq c_i \\ 0 & i > c_i \end{cases},$$

where c_i is the number of input channels. In this way, the input is "copied" into the output and the our model produces the exact same descriptor as an MVCNN with late pooling after the last layer.

C. Implementation details

We include details about our triplet loss implementation and about our procedure for visualization of discrete rotation groups and their homogeneous spaces.

C.1. Triplet loss

We implement a simple triplet loss. During training, we keep a set containing the descriptors for the last seen instance of each class, $Z = \{z_i\}$, where i is the class label. For each entry in the mini-batch, let c be the class and z its descriptor. We take the descriptor in Z of the same class as a positive example (z_c), and chose the hardest between all the others in the set as the negative: $z_n = \text{argmin}_{z_i \in Z,\, i \neq c}(d(z_i, z))$, where d is a distance function. The contribution of this entry to the loss is then,

$$\mathcal{L} = \max(d(z, z_c) - d(z, z_n) + \alpha, 0), \qquad (11)$$

where α is a margin. We use $\alpha = 0.2$ and d is the cosine distance. Note that this method is only used in the "Ours-R" variations of our method.

C.2. Feature map visualization

Our features are functions on a subgroup of the rotation group $\mathbf{SO}(3)$. Since $\mathbf{SO}(3)$ is a 3-manifold (which can be embedded in $\mathbb{R}^5$), visualization is challenging. As we operate on the discrete subgroup of 60 rotations, we choose a solid with icosahedral symmetry and 60 faces as a proxy for visualization – the pentakis dodecahedron, which is the dual of the truncated icosahedron (the "soccer ball" with 60 vertices).

We associate the color of each face with the feature vector at that element of the group. Since the vector is high-dimensional (usually 256 or 512-D), we use PCA over all feature vectors in a layer (or groups of channels in a layer) and project it into the 3 principal components that can be associated with an RGB value. The same idea is applied to visualize functions on the homogeneous spaces, where the dodecahedron and icosahedron are used as proxies.

D. Visualization

We visualize the icosahedral group properties and include more examples of our equivariant feature maps.

D.1. Icosahedral group

The icosahedral group $\mathcal{I}$ is the group of symmetries of the icosahedron, which consists of 60 rotations, as visualized in Figure 8. We show how the equivariance manifests as permutation when rendering multiple views of an object according to the group structure in Figure 7. Figure 9 shows the Cayley Table for $\mathcal{I}$; note that the color assigned for each group element matches the color in Figure 7.

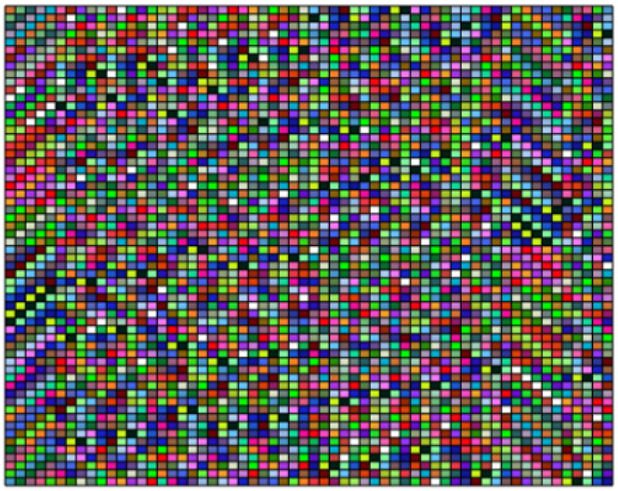

Figure 9: Cayley table for the icosahedral group $\mathcal{I}$. We can see that $\mathcal{I}$ is non-abelian, since the table is not symmetric.

Figure 7: Equivariance of view configurations to $\mathcal{I}$. The views on the left and right are obtained from 3D shapes separated by a 72 deg rotation in the discrete group. We mark corresponding views before and after rotation with same border color. Notice the five first views in the second row – the axis of rotation is aligned with their optical axis; the rotation effect is a shift right of one position for these views. It is clear that when $g \in \mathcal{I}$ is applied to the object, the views are correspondingly permuted in the order given by the Cayley table, showing that the mapping from 3D shape to view set is equivariant.

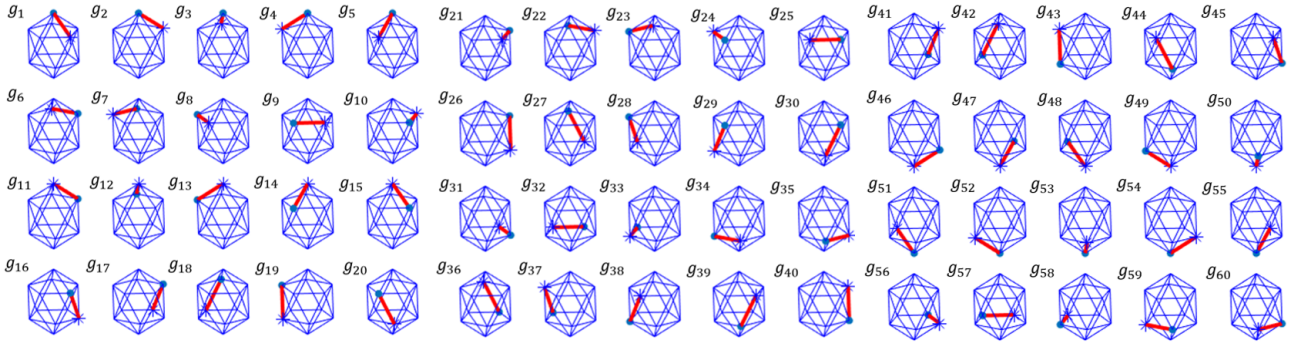

Figure 8: The 60 rotations of the icosahedral group $\mathcal{I}$. We consider g_1 the identity, highlight one edge, and show how each $g_i \in \mathcal{I}$ transforms the highlighted edge.

D.2. Feature maps

We visualize more examples of our equivariant feature maps in Figures 10, 11, 12. Each figure shows 8 different input rotations, the first 5 are from a subgroup of rotations around one axis with 72 deg spacing, the other 3 are from other subgroup with 120 deg spacing. We show the axis of rotation in red. The first column is a view of the input, the second is the initial representation on the group or H-space, and the other 3 are features on each G-CNN layer.

Our method is equivariant to the 60-element discrete rotation group even with only 12 or 20 input views. In Figure 10 we take only 12 input views, giving initial features on the H-space represented by faces of the dodecahedron. Note that the 5 first rotations in this case are in-plane for the views corresponding to the axis of rotation. Due to our procedure described in Section 4.3, this gives an invariant descriptor which can be visualized as the face with constant color. Similarly, in Figure 11, we take 20 views and the invariant descriptor can be seen in the last 3 rotations.

Equivariance is easily visualized on faces neighboring the axis of rotation. For the dodecahedron, we can see cycles of 5 when the axis is on one face and cycles of 3 when the axis is on one vertex. For the icosahedron, we can see cycles of 3 when the axis is on one face and cycles of 5 when the axis is on one vertex. For the pentakis dodecahedron (Figure 12), we can see groups of 5 cells that shift one position when rotation is of 72 deg and groups of 6 cells that shift two positions when rotation is of 120 deg.

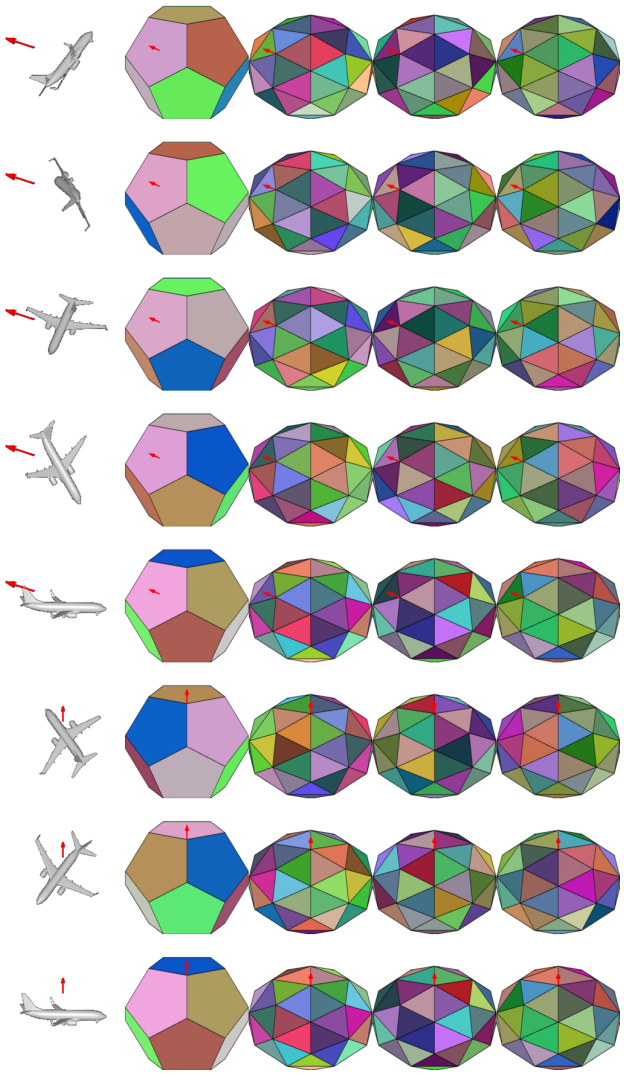

Figure 10: Feature maps with 12 input views. See `animation12.gif` for animated version.

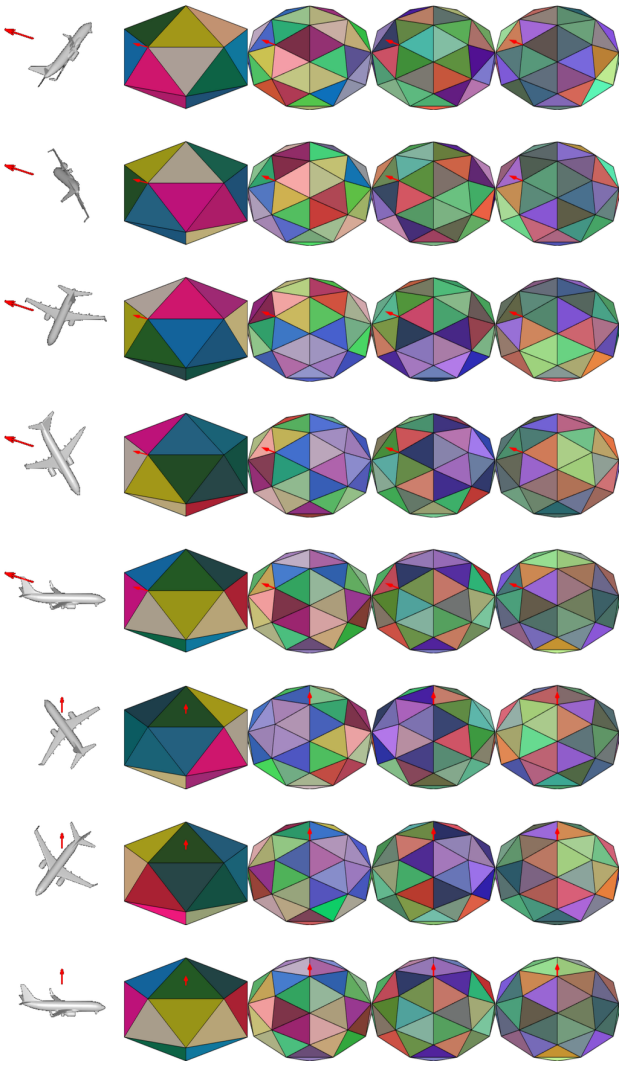

Figure 11: Feature maps with 20 input views. See `animation20.gif` for animated version.

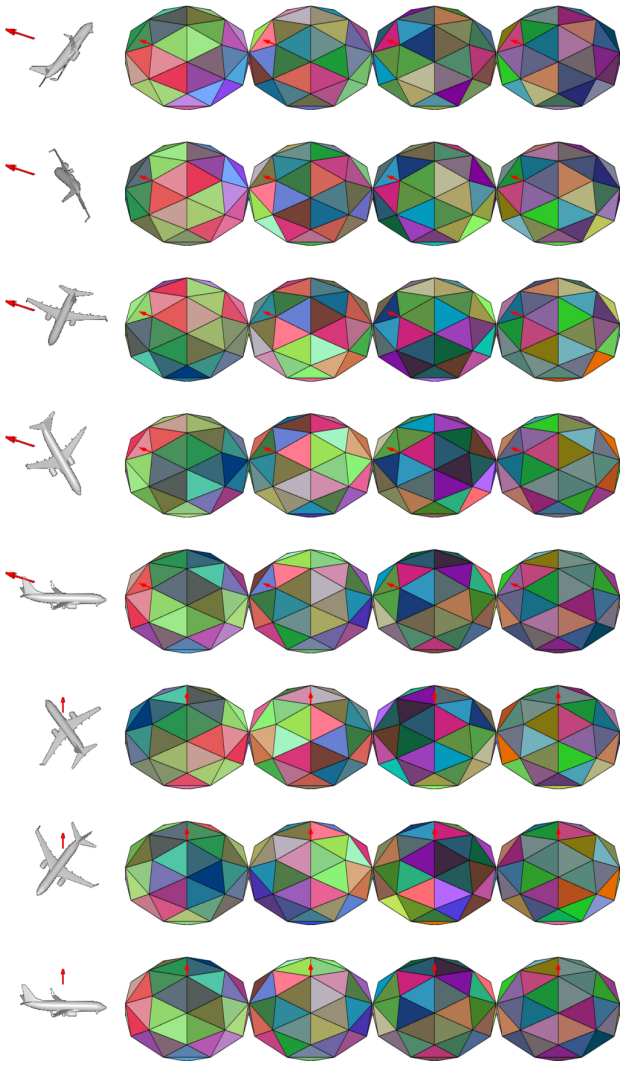

Figure 12: Feature maps with 60 input views. See `animation60.gif` for animated version.