Cross-Domain 3D Equivariant Image Embeddings

Google Research

Carlos Esteves[1,2] machc@seas.upenn.edu
Avneesh Sud[1] asud@google.com
Zhengyi Luo[2] zhengyil@seas.upenn.edu
Kostas Daniilidis[2] kostas@cis.upenn.edu
Ameesh Makadia[1] makadia@google.com

[1]Google Research [2]University of Pennsylvania

Penn Engineering GRASP LABORATORY General Robotics, Automation, Sensing & Perception Lab

Abstract

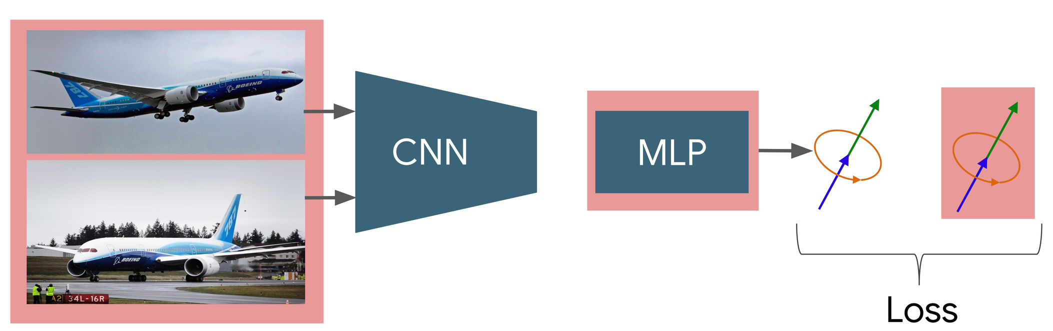
Learning 2D-image embeddings that are *equivariant* to 3D object rotations.

- Our embeddings
  - enable 3D geometric reasoning from 2D inputs
  - generalize to multiple tasks, including pose estimation and novel view synthesis
- Advantages of our approach:
  - reduced sample complexity (by avoiding training on pairs)
  - no task-specific supervision (e.g. no regression or supervision of pose)
  - training only requires a categorized collection of unaligned 3D meshes.

Conventional Approaches

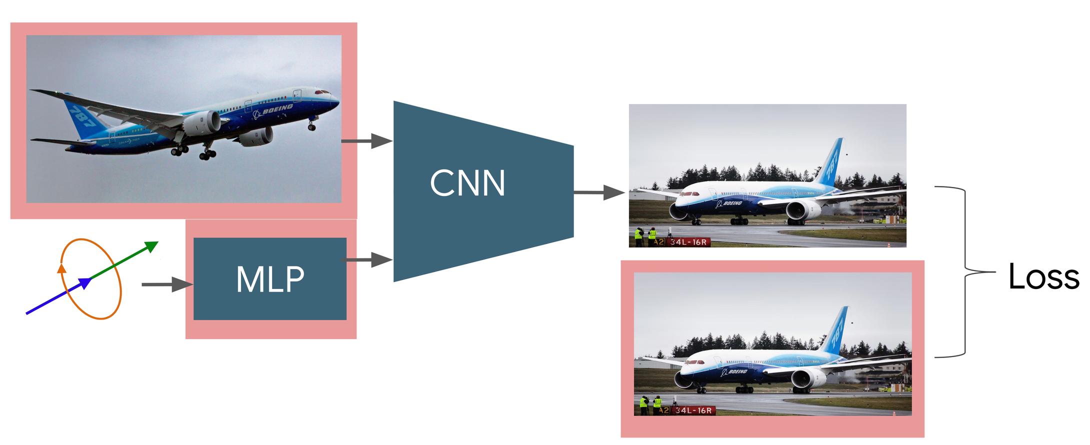
Relative pose estimation

- Need ground truth pose
- Pose regression (tricky)
- Train on pairs of inputs
  - high sample complexity



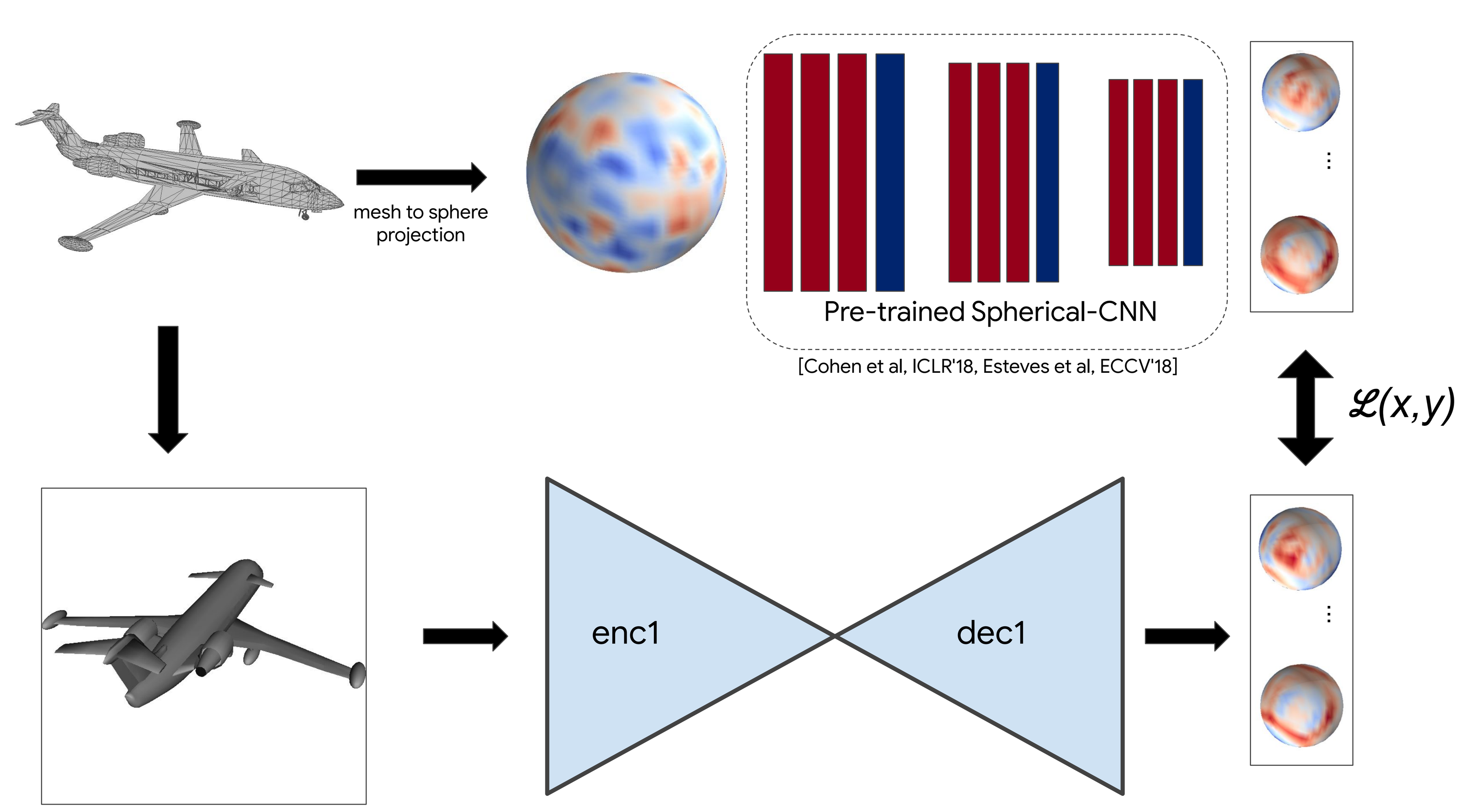
Novel view synthesis

- Pose embedding (tricky)
- Train on input/target pairs
  - high sample complexity



3D Equivariant Embeddings

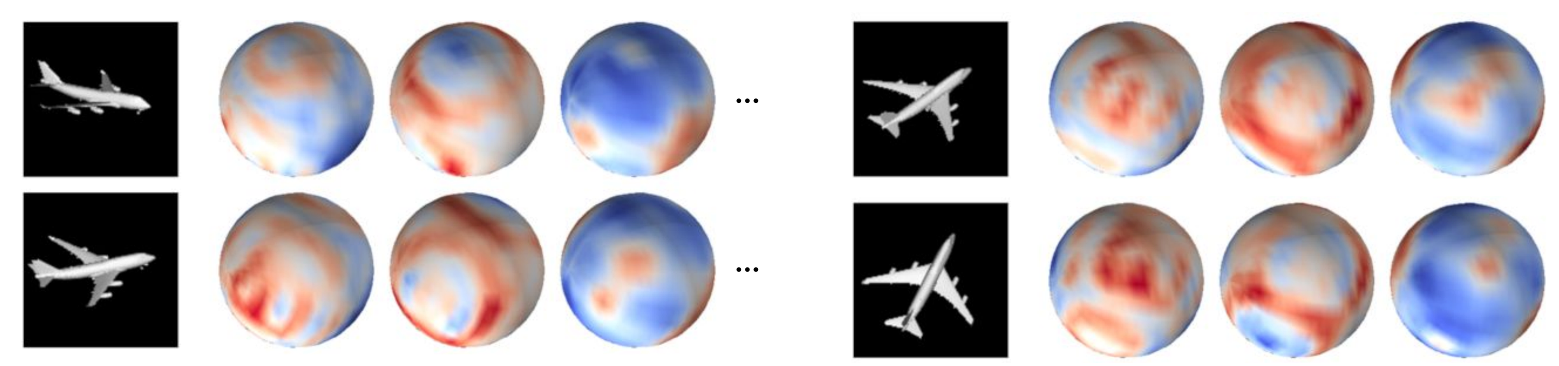
- Our embeddings are **high-dimensional, spherical** functions
- Mapping a 2D image (Euclidean space) to the sphere requires a novel architecture and robust losses
- Supervision from a pre-trained **Spherical CNN (3D rotation equivariant by design)**
- The model produces a 3D equivariant embedding from a **single image**



3D Equivariant Embeddings (details)

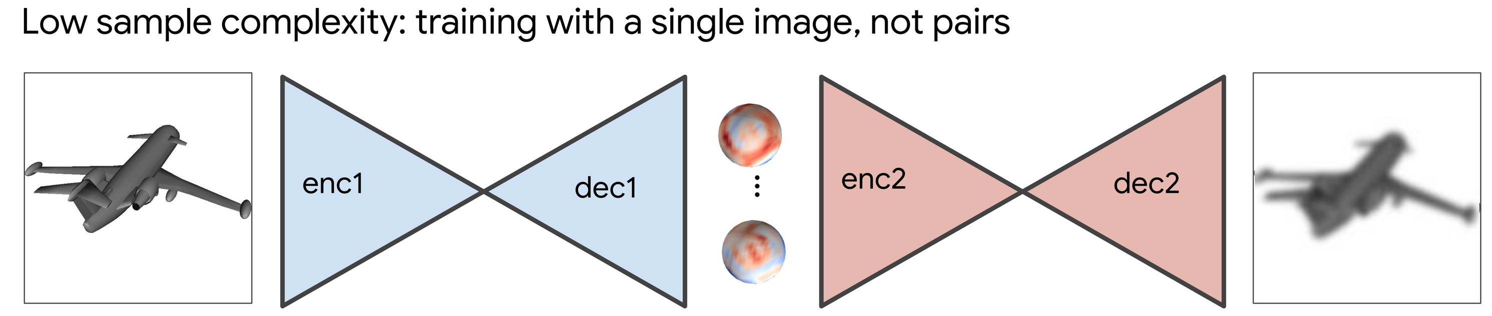
Cross-domain 3D equivariant image embeddings are obtained with

- fully convolutional encoder-decoder inspired by DCGAN (Radford et al, ICLR'16)
- decoder uses equirectangular projection, spherical padding
- Huber loss with weights to handle equirectangular distortions
- skip connections such as in Hourglass (Newell et al, ECCV'16) are avoided for being harmful when crossing domains
- supervising Spherical CNN (Esteves et al, ECCV'18) is trained only once for classification on ModelNet40; we show the obtained embeddings generalize to multiple tasks and datasets.

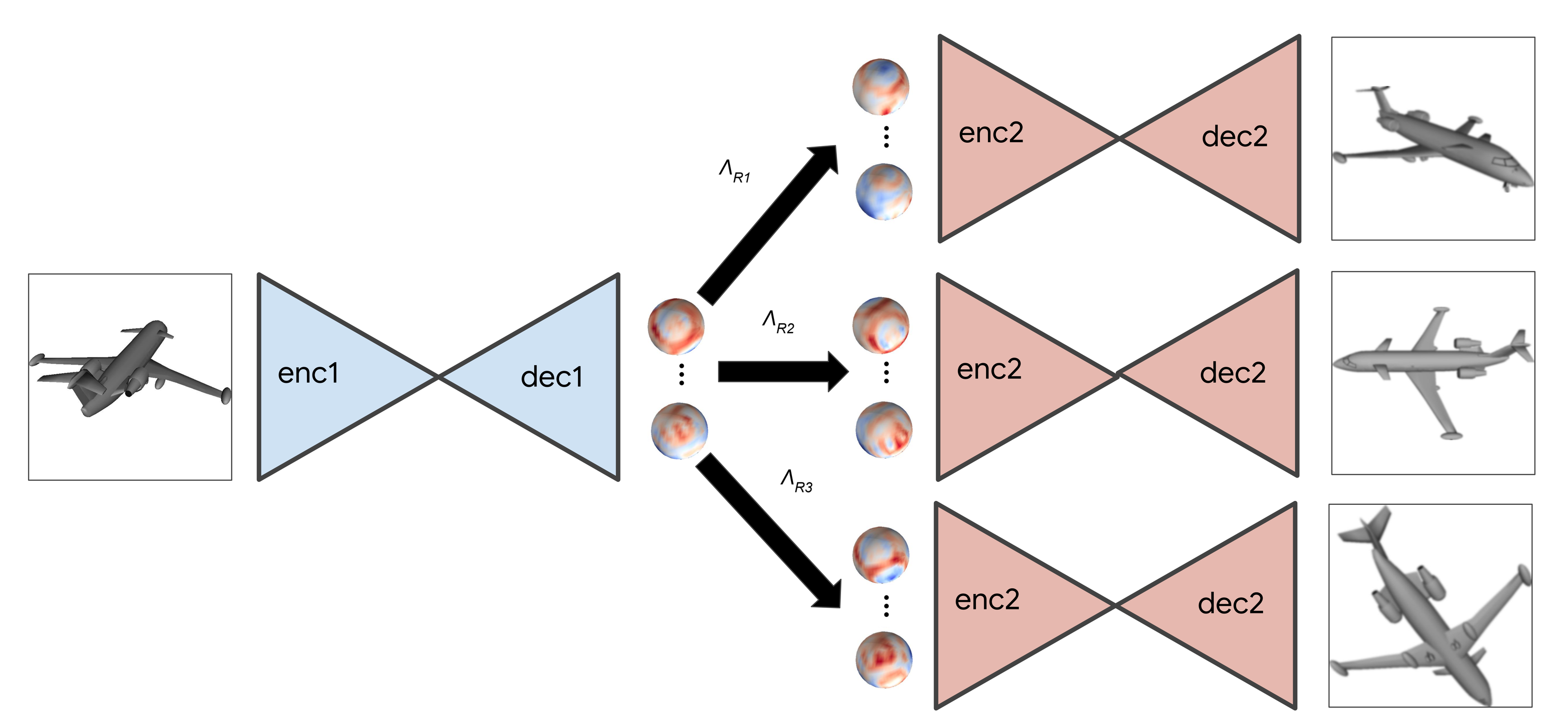


Novel View Synthesis

- We train another network to invert the embeddings with a loss to reconstruct the input
  - architecture similar to a flipped embedding network, with L_2 loss
- Low sample complexity: training with a single image, not pairs



- At test time we embed, rotate, and invert to generate novel views
- No need for pose embeddings (no MLP) or to choose a pose representation



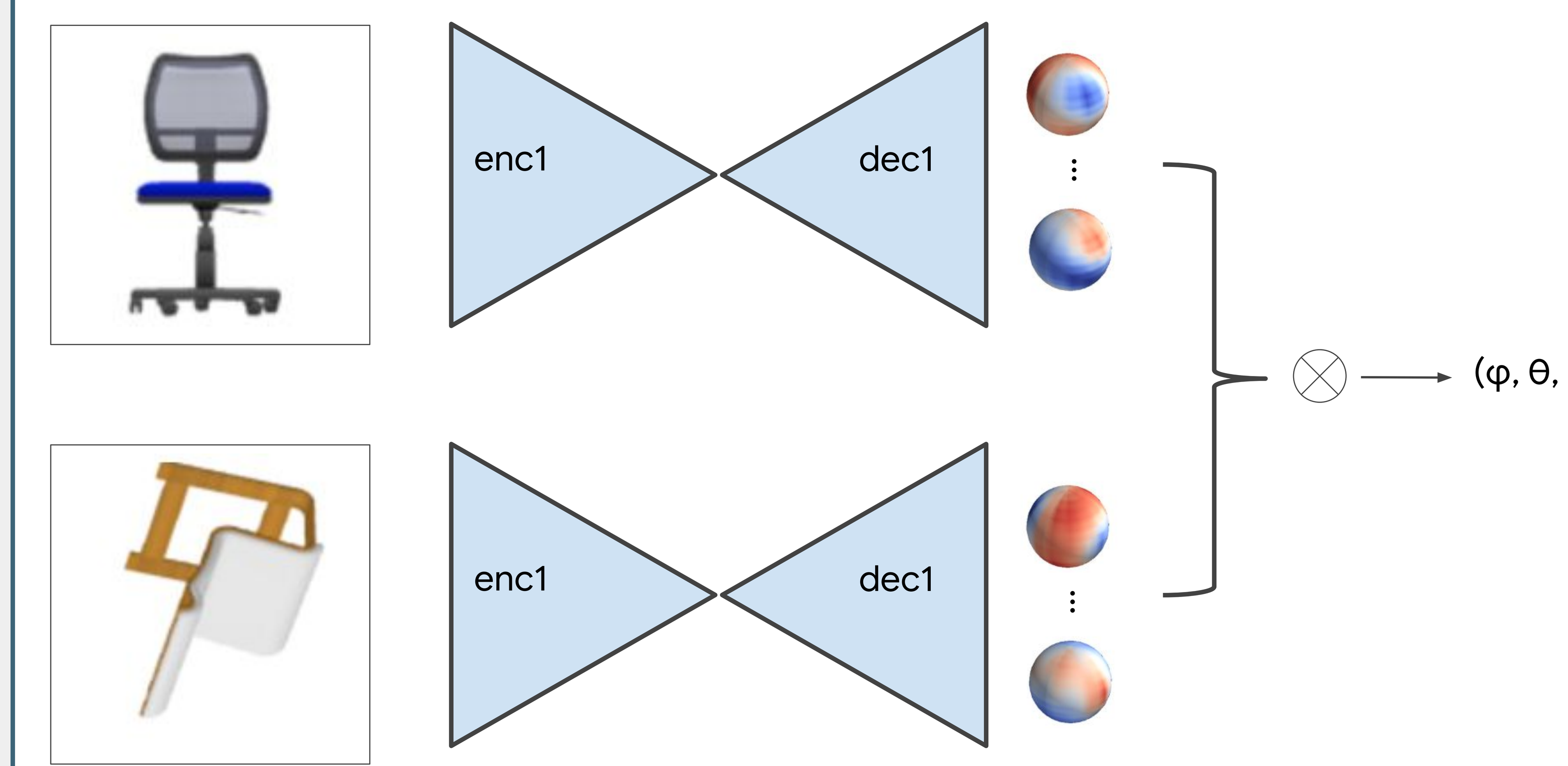
We can generate any novel view from any given view.

Relative Pose Estimation

- Estimate relative pose by maximizing correlation of spherical embeddings:

$$\arg\max_{R\in \mathbf{SO}(3)} G(R) = \sum_{k=0}^{K-1} \int_{p\in S^2} f(y_1)_k(p)\cdot f(y_2)_k(R^T p)\,dp$$

- No direct pose regression (e.g. spatial transformers), no pose supervision
- Can also be applied to image-mesh alignment



Results on ShapeNet shown by rotating one input into another based on estimated relative pose.

Experiments on ShapeNet consider same-instance (SI), inter-instance (II), and 2- and 3-DOF relative pose. Metrics are median error in degrees, and accuracy at 15 and 30 degrees.

			airplane			car			chair			sofa		
			med.	a@15	a@30	med.	a@15	a@30	med.	a@15	a@30	med.	a@15	a@30
2DOF	SI	Ours	**5.17**	**85.3**	**91.9**	**3.70**	**92.2**	92.5	**5.07**	**90.6**	**94.1**	**4.59**	**93.6**	**95.2**
		Regr.	16.9	46.3	68.7	6.55	83.5	**93.1**	13.7	53.9	78.3	17.3	43.2	69.4
		KpNet	6.95	79.4	91.5	div.	div.	div.	6.34	84.7	91.8	9.20	71.3	85.4
	II	Ours	**6.24**	79.0	88.2	**4.73**	73.2	73.3	12.1	59.3	74.4	**10.8**	**58.7**	70.5
		Regr.	20.6	38.7	63.7	7.06	**82.4**	**92.5**	16.8	43.7	72.0	19.6	37.8	66.5
		KpNet	9.07	**79.4**	**91.5**	div.	div.	div.	**8.07**	**79.5**	**90.2**	15.1	49.8	**71.8**
3DOF	SI	Ours	**6.64**	**80.9**	**91.9**	**3.84**	**97.3**	**98.8**	**5.55**	**89.1**	**95.7**	**5.21**	**90.4**	**94.8**
		Regr.	45.4	12.6	31.3	9.83	69.0	86.5	21.7	31.3	64.3	22.2	34.8	61.4
		KpNet	14.9	50.3	76.6	9.12	70.4	80.9	10.8	66.7	85.3	25.0	27.4	57.3
	II	Ours	**7.27**	**76.4**	**89.4**	**4.59**	**92.1**	**93.3**	**12.3**	**59.5**	77.3	**9.66**	**63.9**	**76.0**
		Regr.	44.4	14.1	32.1	10.5	66.5	85.6	25.6	25.1	57.2	24.5	30.9	58.1
		KpNet	16.3	46.0	75.0	10.7	64.4	77.6	13.6	55.4	**81.6**	37.4	12.7	39.8

KeypointNet: Suwajanakorn et al, NIPS'18. Regression: Mahendran et al, CVPR-W'17.

Real images from ObjectNet3D. Median error: 13.75 deg (ours), 36.52 deg (regression).

Conclusion

Geometric image embeddings generalize to a variety of tasks including relative pose estimation and novel view synthesis

Our method for 3D equivariant embeddings:

- avoids difficulties of traditional approaches, (e.g. task-specific supervision, pose embeddings, pose regression)
- requires only aligned image-mesh pairs at training (no alignment across meshes)