Learning image representations equivariant to ego-motion (Supplementary material)

Dinesh Jayaraman
UT Austin
dineshj@cs.utexas.edu

Kristen Grauman
UT Austin
grauman@cs.utexas.edu

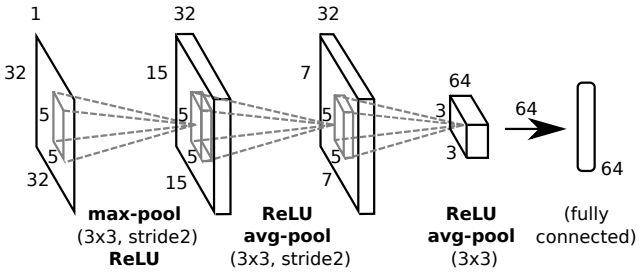

Figure 2. KITTI $\mathbf{z}_{\boldsymbol{\theta}}$ architecture producing D =64-dim. features: 3 convolution layers and a fully connected feature layer (non-linear operations specified along the bottom).

1. KITTI and SUN dataset samples

Some sample images from KITTI and SUN are shown in Fig 1. As they show, these datasets have substantial domain differences. In KITTI, the camera faces the road and has a fixed field of view and camera pitch, and the content is entirely street scenes around Karlsruhe. In SUN, the images are downloaded from the internet, and belong to 397 diverse indoor and outdoor scene categories—most of which have nothing to do with roads.

2. Optimization and hyperparameter selection (Main Sec 4.1)

(Elaborating on para titled "Network architectures and Optimization" 4.1) As mentioned in the paper, for KITTI, we closely follow the cuda-convnet [1] recommended CIFAR-10 architecture: 32 conv(5x5)-max(3x3)-ReLU $\rightarrow$ 32 conv(5x5)-ReLU-avg(3x3) $\rightarrow$ 64 conv(5x5)-ReLU-avg(3x3) $\rightarrow$ D =64 full feature units. A schematic representation for this architecture is shown in Fig 2.

We use Nesterov-accelerated stochastic gradient descent as implemented in Caffe [5], starting from weights randomly initialized according to [3]. The base learning rate and regularization λs are selected with greedy cross-validation. Specifically, for each task, the optimal base learning rate (from 0.1, 0.01, 0.001, 0.0001) was identified for CLSNET. Next, with this base learning rate fixed, the optimal regularizer weight (for DRLIM, TEMPORAL and EQUIV) was selected from a logarithmic grid (steps of $10^{0.5}$). For EQUIV+DRLIM, the DRLIM loss regularizer weight fixed for DRLIM was retained, and only the EQUIV loss weight was cross-validated. The contrastive loss margin parameter δ in Eq (6) in DRLIM, TEMPORAL and EQUIV were set uniformly to 1.0. Since no other part of these objectives (including the softmax classification loss) depends on the scale of features,[1] different choices of margins δ in these methods lead to objective functions with equivalent optima - the features are only scaled by a factor. For EQUIV+DRLIM, we set the DRLIM and EQUIV margins respectively to 1.0 and 0.1 to reflect the fact that the equivariance maps M_g of Eq (5) applied to the representation $\mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})$ of the transformed image must bring it closer to the original image representation $\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x})$ than it was before *i.e.* $\|M_g\mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x}) - \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2 < \|\mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x}) - \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2$.

In addition, to allow fast and thorough experimentation, we set the number of training epochs for each method on each dataset based on a number of initial runs to assess the scale of time usually taken before the classification softmax loss on validation data began to rise *i.e.* overfitting began. All future runs for that method on that data were run to roughly match (to the nearest 5000) the number of epochs identified above. For most cases, this number was of the order of 50000. Batch sizes (for both the classification stack and the Siamese networks) were set to 16 (found to have no major difference from 4 or 64) for NORB-NORB and KITTI-KITTI, and to 128 (selected from 4, 16, 64, 128) for KITTI-SUN, where we found it necessary to increase batch size so that meaningful classification loss gradients were computed in each SGD iteration, and training loss began to fall, despite the large number (397) of classes.

On a single Tesla K-40 GPU machine, NORB-NORB training tasks took $\approx$15 minutes, KITTI-KITTI tasks took $\approx$30 minutes, and KITTI-SUN tasks took $\approx$2 hours.

[1] Technically, the EQUIV objective in Eq (5) may benefit from setting different margins corresponding to the different ego-motion patterns, but we overlook this in favor of scalability and fewer hyperparameters.

Figure 1. (top) Figure from [2] showcasing images from the 4 KITTI location classes (shown here in color; we use grayscale images), and (bottom) Figure from [8] showcasing images from a subset of the 397 SUN classes (shown here in color; see text in main paper for image pre-processing details).

3. Equivariance measurement (Main Sec 4.2)

Computing ρ_g - details In Sec 4.2 in the main paper, we proposed the following measure for equivariance. For each ego-motion g, we measure equivariance separately through the normalized error ρ_g:

$$\rho_g = E\left[\frac{\|\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) - M'_g\mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})\|_2}{\|\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})\|_2}\right], \quad (1)$$

where $E[.]$ denotes the empirical mean, M'_g is the equivariance map, and $\rho_g = 0$ would signify perfect equivariance. We closely follow the equivariance evaluation approach of [6] to solve for the equivariance maps of features produced by each compared method on held-out validation data (cf. Sec 4.1 from the paper), before computing ρ_g. Such maps are produced explicitly by our method, but not the baselines. Thus, as in [6], we compute their maps[2] by solving a least squares minimization problem based on the definition of equivariance in Eq (2) in the paper:

$$M'_g = \arg\min_M \sum_{m(\boldsymbol{y}_i,\boldsymbol{y}_j)=g} \|\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - M\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\|_2. \quad (2)$$

M'_g's computed as above are used to compute ρ_g's as in Eq (1). M'_g and ρ_g are computed on disjoint subsets of the validation data. Since the output features are relatively low in dimension ($D = 100$), we find regularization for Eq (2) unnecessary.

Equivariance results - details While results in the main paper (Table 1) were reported as averages over atomic and composite motions, we present here the results for individual motions in Table 1. While relative trends among the methods remain the same as for the averages reported in the main paper, the new numbers help demonstrate that ρ_g for composite motions is no bigger than for atomic motions, as we would expect from the argument presented in Sec 3.2 in the main paper.

To see this, observe that even among the atomic motions, ρ_g for all methods is lower on the small "up" atomic ego-motion (5°) than it is for the larger "right" ego-motion (20°). Further, the errors for "right" are close to those for the composite motions ("up+right", "up+left" and "down+right"), establishing that while equivariance is diminished for larger motions, it is not affected by whether the motions were used in training or not. In other words, if trained for equivariance to a suitable discrete set of atomic ego-motions (cf. Sec 3.2 in the paper), the feature space generalizes well to new ego-motions.

[2]For uniformity, we do the same recovery of M'_g for our method; our results are similar either way.

Tasks →	atomic		composite		
Datasets ↓	"up (u)"	"right (r)"	"u+r"	"u+l"	"d+r"
random	1.0000	1.0000	1.0000	1.0000	1.0000
CLSNET	0.9276	0.9202	0.9222	0.9138	0.9074
TEMPORAL [7]	0.7140	0.8033	0.8089	0.8061	0.8207
DRLIM [4]	0.5770	0.7038	0.7281	0.7182	0.7325
EQUIV	0.5328	0.6836	0.6913	0.6914	0.7120
EQUIV+DRLIM	**0.5293**	**0.6335**	**0.6450**	**0.6460**	**0.6565**

Table 1. The "normalized error" equivariance measure ρ_g for individual ego-motions (Eq (1)) on NORB, organized as "atomic" (motions in the EQUIV training set) and "composite" (novel) ego-motions.

4. Recognition results (Main Sec 4.3)

Restricted slowness is a weak prior We now present evidence supporting our claim in the paper that the principle of slowness, which penalizes feature variation within small temporal windows, provides a prior that is rather weak. In every stochastic gradient descent (SGD) training iteration for the DRLIM and TEMPORAL networks, we also computed a "slowness" measure that is independent of feature scaling (unlike the DRLIM and TEMPORAL losses of Eq 7 themselves), to better understand the shortcomings of these methods.

Given training pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ annotated as neighbors or non-neighbors by $n_{ij} = 1(|t_i - t_j| \leq T)$ (cf. Eq (7) in the paper), we computed pairwise distances $\Delta_{ij} = d(\mathbf{z}_{\boldsymbol{\theta}(s)}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}(s)}(\boldsymbol{x}_j))$, where $\boldsymbol{\theta}(s)$ is the parameter vector at SGD training iteration s, and $d(.,.)$ is set to the ℓ_2 distance for DRLIM and to the ℓ_1 distance for TEMPORAL (cf. Sec 4).

We then measured how well these pairwise distances Δ_{ij} predict the temporal neighborhood annotation n_{ij}, by measuring the Area Under Receiver Operating Characteristic (AUROC) when varying a threshold on Δ_{ij}.

These "slowness AUROC"s are plotted as a function of training iteration number in Fig 3, for DRLIM and COHERENCE networks trained on the KITTI-SUN task. Compared to the standard random AUROC value of 0.5, these slowness AUROCs tend to be near 0.9 already even before optimization begins, and reach peak AUROCs very close to 1.0 on both training and testing data within about 4000 iterations (batch size 128). This points to a possible weakness in these methods—even with parameters (temporal neighborhood size, regularization λ) cross-validated for recognition, the slowness prior is too weak to regularize feature learning effectively, since strengthening it causes loss of discriminative information. In contrast, our method requires *systematic* feature space responses to ego-motions, and offers a stronger prior.

5. Next-best view selection (Main Sec 4.4)

We now describe our method for next-best view selection for recognition on NORB. Given one view of a NORB ob-

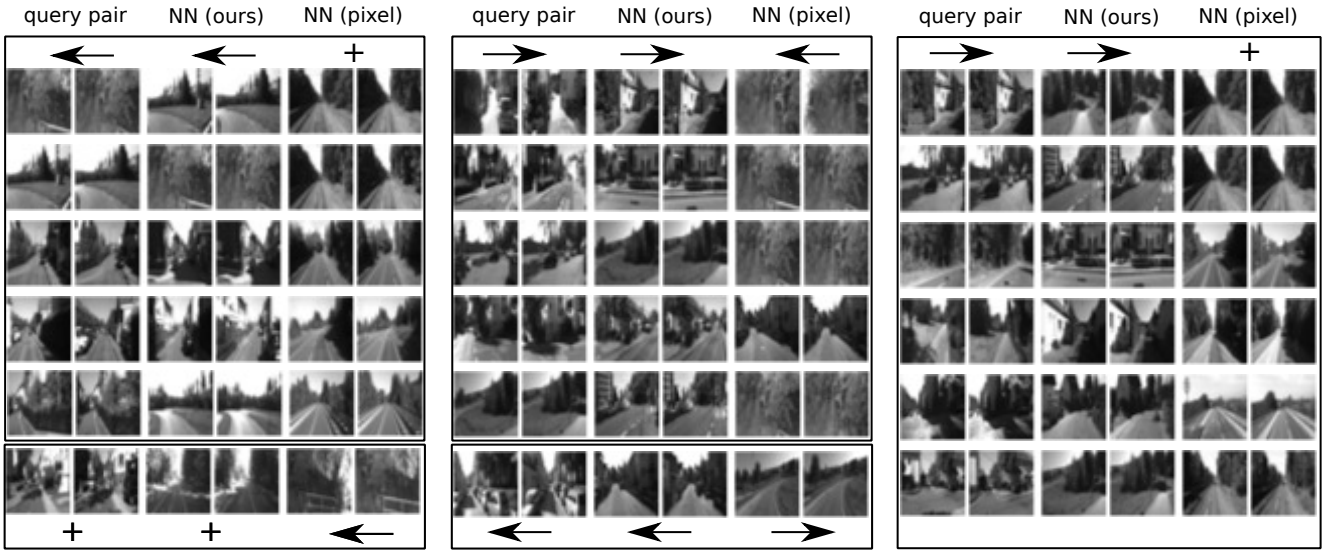

Figure 4. (Contd. from Fig 4) More examples of nearest neighbor image pairs (cols 3 and 4 in each block) in pairwise equivariant feature difference space for various query image pairs (cols 1 and 2 per block). For comparison, cols 5 and 6 show pixel-wise difference-based neighbor pairs. The direction of ego-motion in query and neighbor pairs (inferred from ego-pose vector differences) is indicated above each block.

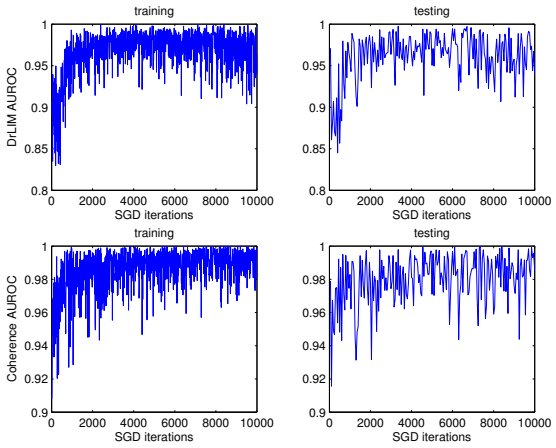

Figure 3. Slowness AUROC on training (left) and testing (right) data for (top) DRLIM (bottom) COHERENCE, showing the weakness of slowness prior.

ject, the task is to tell a hypothetical robot how to move next to help recognize the object *i.e.* which neighboring view would best reduce object prediction uncertainty. We exploit the fact that equivariant features behave predictably under ego-motions to identify the optimal next view.

We limit the choice of next view g to { "up", "down", "up+right" and "up+left" } for simplicity in this preliminary test. We build a k-nearest neighbor (k-NN) image-pair classifier for each possible g, using only training image pairs $(\boldsymbol{x}, g\boldsymbol{x})$ related by the ego-motion g. This classifier C_g takes as input a vector of length $2D$, formed by appending the features of the image pair (each image's representation is of length D) and produces the output probability of each class. So, $C_g([\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}),\ \mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})])$ returns class likelihood probabilities for all 25 NORB classes. Output class probabilities for the k-NN classifier are computed from the histogram of class votes from the k nearest neighbors. We set $k = 25$.

At test time, we first compute features $\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_0)$ on the given starting image $\boldsymbol{x}_0$. Next we predict the feature $\mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x}_0)$ corresponding to each possible surrounding view g, as $M'_g\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_0)$, per the definition of equivariance (cf. Eq 2 in the paper).[3]

With these predicted transformed image features and the pair-wise nearest neighbor class probabilities $C_g(.)$, we may now pick the next-best view as:

$$g^* = \arg\min_g H(C_g([\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_0),\ M'_g\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_0)])), \qquad (3)$$

where $H(.)$ is the information-theoretical entropy function. This selects the view that would produce the least predicted image pair class prediction uncertainty.

6. Qualitative analysis (Main Sec 4.5)

To qualitatively evaluate the impact of equivariant feature learning, we pose a pair-wise nearest neighbor task in the *feature difference* space to retrieve image pairs related by similar ego-motion to a query image pair (details in Supp). Given a learned feature space $\mathbf{z}(.)$ and a query

[3] Equivariance maps M'_g for all methods are computed as described in Sec 3 in this document (and Sec 4.2 in the main paper)

image pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, we form the pairwise feature difference $\boldsymbol{d}_{ij} = \mathbf{z}(\boldsymbol{x}_i) - \mathbf{z}(\boldsymbol{x}_j)$. In an equivariant feature space, other image pairs $(\boldsymbol{x}_k, \boldsymbol{x}_l)$ with similar feature difference vectors $\boldsymbol{d}_{kl} \approx \boldsymbol{d}_{ij}$ would be likely to be related by similar ego-motion to the query pair.[4] This can also be viewed as an analogy completion task, $\boldsymbol{x}_i : \boldsymbol{x}_j = \boldsymbol{x}_k :?$, where the right answer should apply p_{ij} to $\boldsymbol{x}_k$ to obtain $\boldsymbol{x}_l$. For the results in the paper, the closest pair to the query in the learned equivariant feature space is compared to that in the pixel space. Some more examples are shown in Fig 4.

References

[1] Cuda-convnet. https://code.google.com/p/cuda-convnet/. 1

[2] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets Robotics: The KITTI Dataset. *IJRR*, 2013. 2

[3] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *AISTATS*, 2010. 1

[4] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. *CVPR*, 2006. 3

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv*, 2014. 1

[6] K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. *CVPR*, 2015. 3

[7] H. Mobahi, R. Collobert, and J. Weston. Deep Learning from Temporal Coherence in Video. *ICML*, 2009. 3

[8] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *CVPR*, 2010. 2

[4] Note that in our model of equivariance, this isn't strictly true, since the pair-wise difference vector $M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x})$ need not actually be fixed for a given transformation g, $\forall \boldsymbol{x}$. For small motions (and the right kinds of equivariant maps M_g), this still holds approximately, as we find in practice.