In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

Learning image representations tied to ego-motion

Dinesh Jayaraman
The University of Texas at Austin
dineshj@cs.utexas.edu

Kristen Grauman
The University of Texas at Austin
grauman@cs.utexas.edu

Abstract

Understanding how images of objects and scenes behave in response to specific ego-motions is a crucial aspect of proper visual development, yet existing visual learning methods are conspicuously disconnected from the physical source of their images. We propose to exploit proprioceptive motor signals to provide unsupervised regularization in convolutional neural networks to learn visual representations from egocentric video. Specifically, we enforce that our learned features exhibit equivariance i.e. *they respond predictably to transformations associated with distinct ego-motions. With three datasets, we show that our unsupervised feature learning approach significantly outperforms previous approaches on visual recognition and next-best-view prediction tasks. In the most challenging test, we show that features learned from video captured on an autonomous driving platform improve large-scale scene recognition in static images from a disjoint domain.*

1. Introduction

How is visual learning shaped by ego-motion? In their famous "kitten carousel" experiment, psychologists Held and Hein examined this question in 1963 [10]. To analyze the role of self-produced movement in perceptual development, they designed a carousel-like apparatus in which two kittens could be harnessed. For eight weeks after birth, the kittens were kept in a dark environment, except for one hour a day on the carousel. One kitten, the "active" kitten, could move freely of its own volition while attached. The other kitten, the "passive" kitten, was carried along in a basket and could not control his own movement; rather, he was forced to move in exactly the same way as the active kitten. Thus, both kittens received the same visual experience. However, while the active kitten simultaneously experienced signals about his own motor actions, the passive kitten did not. The outcome of the experiment is remarkable. While the active kitten's visual perception was indistinguishable from kittens raised normally, the passive kitten suffered fundamental problems. The implication is clear: proper perceptual development requires leveraging *self-generated movement in concert with visual feedback*.

Figure 2. We learn visual features from egocentric video that respond predictably to observer egomotion.

We contend that today's visual recognition algorithms are crippled much like the passive kitten. The culprit: learning from "bags of images". Ever since statistical learning methods emerged as the dominant paradigm in the recognition literature, the norm has been to treat images as i.i.d. draws from an underlying distribution. Whether learning object categories, scene classes, body poses, or features themselves, the idea is to discover patterns within a collection of snapshots, blind to their physical source. So is the answer to learn from video? Only partially. Without leveraging the accompanying motor signals initiated by the videographer, learning from video data does *not* escape the passive kitten's predicament.

Inspired by this concept, we propose to treat visual learning as an embodied process, where the visual experience is inextricably linked to the motor activity behind it.[1] In particular, our goal is to learn representations that exploit the parallel signals of ego-motion and pixels. We hypothesize that downstream processing will benefit from a feature space that preserves the connection between "how I move" and "how my visual surroundings change".

To this end, we cast the problem in terms of unsupervised equivariant feature learning. During training, the input image sequences are accompanied by a synchronized stream of ego-motor sensor readings; however, they need

[1] Depending on the context, the motor activity could correspond to either the 6-DOF ego-motion of the observer moving in the scene or the second-hand motion of an object being actively manipulated, e.g., by a person or robot's end effectors.

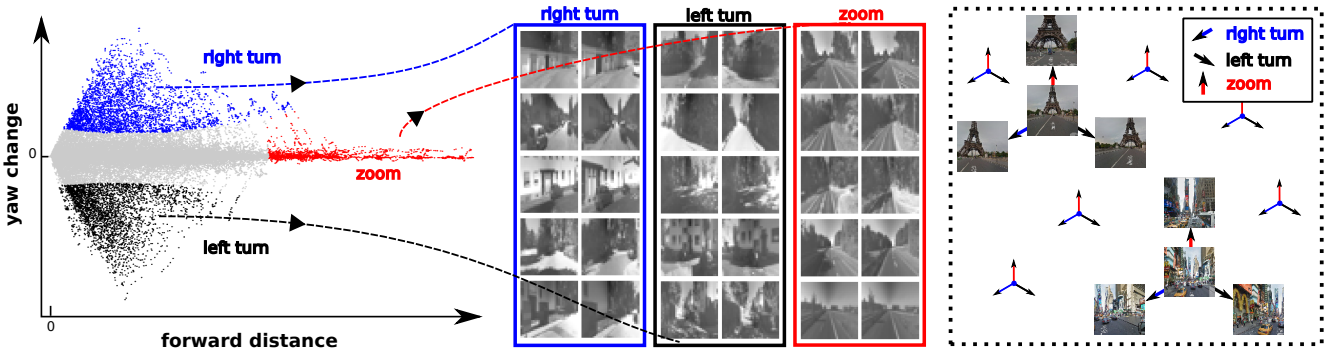

Figure 1. Our goal is to learn a feature space equivariant to ego-motion. We train with image pairs from video accompanied by their sensed ego-poses (left and center), and produce a feature mapping such that two images undergoing the same ego-pose *change* move similarly in the feature space (right). **Left:** Scatter plot of motions $(\boldsymbol{y}_i - \boldsymbol{y}_j)$ among pairs of frames $\leq$ 1s apart in video from KITTI car-mounted camera, clustered into motion patterns p_{ij}. **Center:** Frame pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ from the "right turn", "left turn" and "zoom" motion patterns. **Right:** An illustration of the equivariance property we seek in the learned feature space. Pairs of frames corresponding to each ego-motion pattern ought to have predictable relative positions in the learned feature space. Best seen in color.

not possess any semantic labels. The ego-motor signal could correspond, for example, to the inertial sensor measurements received alongside video on a wearable or car-mounted camera. The objective is to learn a feature mapping from pixels in a video frame to a space that is *equivariant* to various motion classes. In other words, the learned features should *change in predictable and systematic ways as a function of the transformation applied to the original input.* See Fig 1. We develop a convolutional neural network (CNN) approach that optimizes a feature map for the desired egomotion-based equivariance. To exploit the features for recognition, we augment the network with a classification loss when class-labeled images are available. In this way, ego-motion serves as side information to regularize the features learned, which we show facilitates category learning when labeled examples are scarce.

In sharp contrast to our idea, previous work on visual features—whether hand-designed or learned—primarily targets feature *invariance*. Invariance is a special case of equivariance, where transformations of the input have no effect. Typically, one seeks invariance to small transformations, e.g., the orientation binning and pooling operations in SIFT/HOG and modern CNNs both target invariance to local translations and rotations. While a powerful concept, invariant representations require a delicate balance: "too much" invariance leads to a loss of useful information or discriminability. In contrast, more general equivariant representations are intriguing for their capacity to impose structure on the output space without forcing a loss of information. Equivariance is "active" in that it exploits observer motor signals, like Hein and Held's active kitten.

Our main contribution is a novel feature learning approach that couples ego-motor signals and video. To our knowledge, ours is the first attempt to ground feature learning in physical activity. The limited prior work on unsupervised feature learning with video [21, 23, 20, 8] learns only passively from observed scene dynamics, uninformed by explicit motor sensory cues. Furthermore, while equivariance is explored in some recent work, unlike our idea, it typically focuses on 2D image transformations as opposed to 3D ego-motion [13, 25] and considers existing features [29, 16]. Finally, whereas existing methods that learn from image transformations focus on view synthesis applications [11, 14, 20], we explore recognition applications of learning jointly equivariant and discriminative feature maps.

We apply our approach to three public datasets. On pure equivariance as well as recognition tasks, our method consistently outperforms the most related techniques in feature learning. In the most challenging test of our method, we show that features learned from video captured on a vehicle can improve image recognition accuracy on a disjoint domain. In particular, we use unlabeled KITTI [6, 7] car data to regularize feature learning for the 397-class scene recognition task for the SUN dataset [33]. Our results show the promise of departing from the "bag of images" mindset, in favor of an embodied approach to feature learning.

2. Related work

Invariant features Invariance is a special case of equivariance, wherein a transformed output remains identical to its input. Invariance is known to be valuable for visual representations. Descriptors like SIFT, HOG, and aspects of CNNs like pooling and convolution, are hand-designed for invariance to small shifts and rotations. Feature learning work aims to *learn* invariances from data [26, 27, 30, 28, 5]. Strategies include augmenting training data by perturbing image instances with label-preserving transformations [27, 30, 5], and inserting linear transformation operators into the feature learning algorithm [28].

Most relevant to our work are feature learning meth-

ods based on temporal coherence and "slow feature analysis" [31, 9, 21]. The idea is to require that learned features vary slowly over continuous video, since visual stimuli can only gradually change between adjacent frames. Temporal coherence has been explored for unsupervised feature learning with CNNs [21, 36, 8, 3, 18], with applications to dimensionality reduction [9], object recognition [21, 36], and metric learning [8]. Temporal coherence of inferred body poses in unlabeled video is exploited for invariant recognition in [4]. These methods exploit video as a source of free supervision to achieve invariance, analogous to the image perturbations idea above. In contrast, our method exploits video coupled with ego-motor signals to achieve the more general property of equivariance.

Equivariant representations Equivariant features can also be hand-designed or learned. For example, equivariant or "co-variant" operators are designed to detect repeatable interest points [29]. Recent work explores ways to learn descriptors with in-plane translation/rotation equivariance [13, 25]. While the latter does perform feature learning, its equivariance properties are crafted for specific 2D image transformations. In contrast, we target more complex equivariances arising from natural observer motions (3D ego-motion) that cannot easily be crafted, and our method learns them from data.

Methods to learn representations with disentangled latent factors [11, 14] aim to sort properties like pose, illumination *etc*. into distinct portions of the feature space. For example, the transforming auto-encoder learns to explicitly represent instantiation parameters of object parts in equivariant hidden layer units [11]. Such methods target equivariance in the limited sense of inferring pose parameters, which are appended to a conventional feature space designed to be invariant. In contrast, our formulation encourages *equivariance* over the *complete* feature space; we show the impact as an unsupervised regularizer when training a recognition model with limited training data.

The work of [16] quantifies the invariance/equivariance of various standard representations, including CNN features, in terms of their responses to specified in-plane 2D image transformations (affine warps, flips of the image). We adopt the definition of equivariance used in that work, but our goal is entirely different. Whereas [16] quantifies the equivariance of existing descriptors, our approach learns a feature space that is equivariant.

Learning transformations Other methods train with pairs of transformed images and infer an implicit representation for the transformation itself. In [19], bilinear models with multiplicative interactions are used to learn content-independent "motion features" that encode only the transformation between image pairs. One such model, the "gated autoencoder" is extended to perform sequence prediction for video in [20]. Recurrent neural networks combined with a grammar model of scene dynamics can also predict future frames in video [23]. Whereas these methods learn a representation for image pairs (or tuples) related by some transformation, we learn a representation for individual images in which the behavior under transformations is predictable. Furthermore, whereas these prior methods abstract away the image content, our method preserves it, making our features relevant for recognition.

Egocentric vision There is renewed interest in egocentric computer vision methods, though none perform feature learning using motor signals and pixels in concert as we propose. Recent methods use ego-motion cues to separate foreground and background [24, 34] or infer the first-person gaze [35, 17]. While most work relies solely on apparent image motion, the method of [34] exploits a robot's motor signals to detect moving objects and [22] uses reinforcement learning to form robot movement policies by exploiting correlations between motor commands and observed motion cues.

3. Approach

Our goal is to learn an image representation that is equivariant with respect to ego-motion transformations. Let $\boldsymbol{x}_i \in \mathcal{X}$ be an image in the original pixel space, and let $\boldsymbol{y}_i \in \mathcal{Y}$ be its associated ego-pose representation. The ego-pose captures the available motor signals, and could take a variety of forms. For example, $\mathcal{Y}$ may encode the complete observer camera pose (its position in 3D space, pitch, yaw, roll), some subset of those parameters, or any reading from a motor sensor paired with the camera.

As input to our learning algorithm, we have a training set $\mathcal{U}$ of N_u image pairs and their associated ego-poses, $\mathcal{U} = \{\langle(\boldsymbol{x}_i, \boldsymbol{x}_j), (\boldsymbol{y}_i, \boldsymbol{y}_j)\rangle\}_{(i,j)=1}^{N_u}$. The image pairs originate from video sequences, though they need not be adjacent frames in time. The set may contain pairs from multiple videos and cameras. Note that this training data does *not* have any semantic labels (object categories, *etc*.); they are "labeled" only in terms of the ego-motor sensor readings.

In the following, we first explain how to translate ego-pose information into pairwise "motion pattern" annotations (Sec 3.1). Then, Sec 3.2 defines the precise nature of the equivariance we seek, and Sec 3.3 defines our learning objective. Sec 3.4 shows how our equivariant feature learning scheme may be used to enhance recognition with limited training data. Finally, in Sec 3.5, we show how a feedforward neural network architecture may be trained to produce the desired equivariant feature space.

3.1. Mining discrete ego-motion patterns

First we want to organize training sample pairs into a discrete set of ego-motion patterns. For instance, one ego-

motion pattern might correspond to "tilt downwards by approximately 20°". While one could collect new data explicitly controlling for the patterns (e.g., with a turntable and camera rig), we prefer a data-driven approach that can leverage video and ego-pose data collected "in the wild".

To this end, we discover clusters among pose difference vectors $\boldsymbol{y}_i - \boldsymbol{y}_j$ for pairs (i, j) of temporally close frames from video (typically $\lessapprox$1 second apart; see Sec 4.1 for details). For simplicity we apply k-means to find G clusters, though other methods are possible. Let $p_{ij} \in \mathcal{P} = \{1, \ldots, G\}$ denote the motion pattern ID, *i.e.*, the cluster to which $(\boldsymbol{y}_i, \boldsymbol{y}_j)$ belongs. We can now replace the ego-pose vectors in $\mathcal{U}$ with motion pattern IDs: $\langle(\boldsymbol{x}_i, \boldsymbol{x}_j), p_{ij}\rangle$. [2]

The left panel of Fig 1 illustrates a set of motion patterns discovered from videos in the KITTI [6] dataset, which are captured from a moving car. Here $\mathcal{Y}$ consists of the position and yaw angle of the camera. So, we are clustering a 2D space consisting of forward distance and change in yaw. As illustrated in the center panel, the largest clusters correspond to the car's three primary ego-motions: turning left, turning right, and going forward.

3.2. Ego-motion equivariance

Given $\mathcal{U}$, we wish to learn a feature mapping function $\mathbf{z}_{\boldsymbol{\theta}}(.) : \mathcal{X} \rightarrow \mathcal{R}^D$ parameterized by $\boldsymbol{\theta}$ that maps a single image to a D-dimensional vector space that is equivariant to ego-motion. To be equivariant, the function $\mathbf{z}_{\boldsymbol{\theta}}$ must respond *systematically* and *predictably* to ego-motion:

$$\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j) \approx f(\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i, \boldsymbol{y}_j), \tag{1}$$

for some function f. We consider equivariance for linear functions $f(.)$, following [16]. In this case, $\mathbf{z}_{\boldsymbol{\theta}}$ is said to be equivariant with respect to some transformation g if there exists a $D \times D$ matrix[3] M_g such that:

$$\forall \boldsymbol{x} \in \mathcal{X} : \mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x}) \approx M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}). \tag{2}$$

Such an M_g is called the "equivariance map" of g on the feature space $\mathbf{z}_{\boldsymbol{\theta}}(.)$. It represents the affine transformation in the feature space that corresponds to transformation g in the pixel space. For example, suppose a motion pattern g corresponds to a yaw turn of 20°, and $\boldsymbol{x}$ and $g\boldsymbol{x}$ are the images observed before and after the turn, respectively. Equivariance demands that there is some matrix M_g that maps the pre-turn image to the post-turn image, once those images are expressed in the feature space $\mathbf{z}_{\boldsymbol{\theta}}$. Hence, $\mathbf{z}_{\boldsymbol{\theta}}$ "organizes" the feature space in such a way that movement in a particular direction in the feature space (here, as computed by multiplication with M_g) has a predictable outcome. The linear case, as also studied in [16], ensures that the structure of the mapping has a simple form, and is convenient for learning since M_g can be encoded as a fully connected layer in a neural network.

While prior work [13, 25] focuses on equivariance where g is a 2D image warp, we explore the case where $g \in \mathcal{P}$ is an ego-motion pattern (cf. Sec 3.1) reflecting the observer's 3D movement in the world. In theory, appearance changes of an image in response to an observer's ego-motion are not determined by the ego-motion alone. They also depend on the depth map of the scene and the motion of dynamic objects in the scene. One could easily augment either the frames $\boldsymbol{x}_i$ or the ego-pose $\boldsymbol{y}_i$ with depth maps, when available. Non-observer motion appears more difficult, especially in the face of changing occlusions and newly appearing objects. However, our experiments indicate we can learn effective representations even with dynamic objects. In our implementation, we train with pairs relatively close in time, so as to avoid some of these pitfalls.

While during training we target equivariance for the discrete set of G ego-motions, the learned feature space will *not* be limited to preserving equivariance for pairs originating from the same ego-motions. This is because the linear equivariance maps are composable. If we are operating in a space where every ego-motion can be composed as a sequence of "atomic" motions, equivariance to those atomic motions is sufficient to guarantee equivariance to all motions. To see this, suppose that the maps for "turn head right by 10°" (ego-motion pattern r) and "turn head up by 10°" (ego-motion pattern u) are respectively M_r and M_u, *i.e.*, $\mathbf{z}(r\boldsymbol{x}) = M_r\mathbf{z}(\boldsymbol{x})$ and $\mathbf{z}(u\boldsymbol{x}) = M_u\mathbf{z}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$. Now for a novel diagonal motion d that can be composed from these atomic motions as $d = r \circ u$, we have

$$\mathbf{z}(d\boldsymbol{x}) = \mathbf{z}((r \circ u)\boldsymbol{x}) = M_r\mathbf{z}(u\boldsymbol{x}) = M_r M_u \mathbf{z}(\boldsymbol{x}), \tag{3}$$

so that $M_d = M_r M_u$ is the equivariance map for novel ego-motion d, even though d was not among $1, \ldots, G$. This property lets us restrict our attention to a relatively small number of discrete ego-motion patterns during training, and still learn features equivariant w.r.t. new ego-motions.

3.3. Equivariant feature learning objective

We now design a loss function that encourages the learned feature space $\mathbf{z}_{\boldsymbol{\theta}}$ to exhibit equivariance with respect to each ego-motion pattern. Specifically, we would like to learn the optimal feature space parameters $\boldsymbol{\theta}^*$ jointly with its equivariance maps $\mathcal{M}^* = \{M_1^*, \ldots, M_G^*\}$ for the motion pattern clusters 1 through G (cf. Sec 3.1).

To achieve this, a naive translation of the definition of equivariance in Eq (2) into a minimization problem over feature space parameters $\boldsymbol{\theta}$ and the $D \times D$ equivariance map candidate matrices $\mathcal{M}$ would be as follows:

$$(\boldsymbol{\theta}^*, \mathcal{M}^*) = \arg\min_{\boldsymbol{\theta}, \mathcal{M}} \sum_g \sum_{\{(i,j):p_{ij}=g\}} d\left(M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\right), \tag{4}$$

[2] For movement with d degrees of freedom, setting $G \approx d$ should suffice (cf. Sec 3.2). We chose small G for speed and did not vary it.

[3] bias dimension assumed to be included in D for notational simplicity

where $d(.,.)$ is a distance measure. This problem can be decomposed into G independent optimization problems, one for each motion, corresponding only to the inner summation above, and dealing with disjoint data. The g-th such problem requires only that training frame pairs annotated with motion pattern $p_{ij} = g$ approximately satisfy Eq (2).

However, such a formulation admits problematic solutions that perfectly optimize it, *e.g.* for the trivial all-zero feature space $\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{0}, \forall \boldsymbol{x} \in \mathcal{X}$ with M_g set to the all-zeros matrix for all g, the loss above evaluates to zero. To avoid such solutions, and to force the learned M_g's to be different from one another (since we would like the learned representation to respond *differently* to different ego-motions), we simultaneously account for the "negatives" of each motion pattern. Our learning objective is:

$$(\boldsymbol{\theta}^*, \mathcal{M}^*) = \arg\min_{\boldsymbol{\theta}, \mathcal{M}} \sum_{g,i,j} d_g \left(M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j), p_{ij} \right), \tag{5}$$

where $d_g(.,.,.)$ is a "contrastive loss" [9] specific to motion pattern g:

$$d_g(\boldsymbol{a}, \boldsymbol{b}, c) = \mathbb{1}(c = g) d(\boldsymbol{a}, \boldsymbol{b}) + \mathbb{1}(c \neq g) \max(\delta - d(\boldsymbol{a}, \boldsymbol{b}), 0), \tag{6}$$

where $\mathbb{1}(.)$ is the indicator function. This contrastive loss penalizes distance between $\boldsymbol{a}$ and $\boldsymbol{b}$ in "positive" mode (when $c = g$), and pushes apart pairs in "negative" mode (when $c \neq g$), up to a minimum margin distance specified by the constant δ. We use the ℓ_2 norm for the distance $d(.,.)$.

In our objective in Eq (5), the contrastive loss operates in the latent feature space. For pairs belonging to cluster g, the contrastive loss d_g penalizes feature space distance between the first image and its transformed pair, similar to Eq (4) above. For pairs belonging to clusters other than g, d_g requires that the transformation defined by M_g must not bring the image representations close together. In this way, our objective learns the M_g's jointly. It ensures that distinct ego-motions, when applied to an input $\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x})$, map it to different locations in feature space.

We want to highlight the important distinctions between our objective and the "temporal coherence" objective of [21] for slow feature analysis. Written in our notation, the objective of [21] may be stated as:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i,j} d_1(\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j), \mathbb{1}(|t_i - t_j| \leq T)), \tag{7}$$

where t_i, t_j are the video time indices of $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ and T is a temporal neighborhood size hyperparameter. This loss encourages the representations of nearby frames to be similar to one another. However, crucially, it does not account for the nature of the ego-motion between the frames. Accordingly, while temporal coherence helps learn invariance to small image changes, it does not target a (more general) equivariant space. Like the passive kitten from Hein and Held's experiment, the temporal coherence constraint watches video to passively learn a representation; like the active kitten, our method registers the *observer motion* explicitly with the video to learn more effectively, as we will demonstrate in results.

3.4. Regularizing a recognition task

While we have thus far described our formulation for generic equivariant image representation learning, it can optionally be used for visual recognition tasks. Suppose that in addition to the ego-pose annotated pairs $\mathcal{U}$ we are also given a small set of N_l class-labeled static images, $\mathcal{L} = \{(\boldsymbol{x}_k, c_k\}_{k=1}^{N_l}$, where $c_k \in \{1, \ldots, C\}$. Let L_e denote the unsupervised equivariance loss of Eq (5). We can integrate our unsupervised feature learning scheme with the recognition task, by optimizing a misclassification loss together with L_e. Let W be a $D \times C$ matrix of classifier weights. We solve jointly for W and the maps $\mathcal{M}$:

$$(\boldsymbol{\theta}^*, W^*, \mathcal{M}^*) = \arg\min_{\boldsymbol{\theta}, W, \mathcal{M}} L_c(\boldsymbol{\theta}, W, \mathcal{L}) + \lambda L_e(\boldsymbol{\theta}, \mathcal{M}, \mathcal{U}), \tag{8}$$

where L_c denotes the softmax loss over the learned features, $L_c(W, \mathcal{L}) = -\frac{1}{N_l} \sum_{i=1}^{N_l} \log(\sigma_{c_k}(W \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$, and $\sigma_{c_k}(.)$ is the softmax probability of the correct class. The regularizer weight λ is a hyperparameter. Note that neither the supervised training data $\mathcal{L}$ nor the testing data for recognition are required to have any associated sensor data. Thus, our features are applicable to standard image recognition tasks.

In this use case, the unsupervised ego-motion equivariance loss encodes a prior over the feature space that can improve performance on the supervised recognition task with limited training examples. We hypothesize that a feature space that embeds knowledge of how objects change under different viewpoints / manipulations allows a recognition system to, in some sense, hallucinate new views of an object to improve performance.

3.5. Form of the feature mapping function $\mathbf{z}_{\boldsymbol{\theta}}(.)$

For the mapping $\mathbf{z}_{\boldsymbol{\theta}}(.)$, we use a convolutional neural network architecture, so that the parameter vector $\boldsymbol{\theta}$ now represents the layer weights. The loss L_e of Eq (5) is optimized by sharing the weight parameters $\boldsymbol{\theta}$ among two identical stacks of layers in a "Siamese" network [2, 9, 21], as shown in the top two rows of Fig 3. Image pairs from $\mathcal{U}$ are fed into these two stacks. Both stacks are initialized with identical random weights, and identical gradients are passed through them in every training epoch, so that the weights remain tied throughout. Each stack encodes the feature map that we wish to train, $\mathbf{z}_{\boldsymbol{\theta}}$.

To optimize Eq (5), an array of equivarance maps $\mathcal{M}$, each represented by a fully connected layer, is connected to

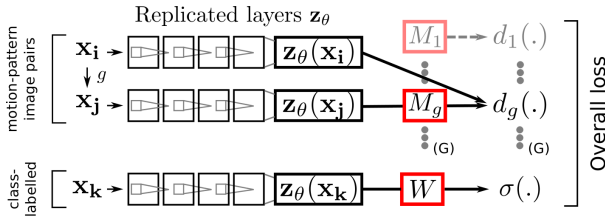

Figure 3. Training setup: (top) "Siamese network" for computing the equivariance loss of Eq (5), together with (bottom) a third tied stack for computing the supervised recognition softmax loss as in Eq (8). See Sec 4.1 and Supp for exact network specifications.

the top of the second stack. Each such equivariance map then feeds into a motion-pattern-specific contrastive loss function d_g, whose other inputs are the first stack output and the ego-motion pattern ID p_{ij}.

To optimize Eq (8), in addition to the Siamese net that minimizes L_e as above, the supervised softmax loss is minimized through a third replica of the $\mathbf{z_\theta}$ layer stack with weights tied to the two Siamese networks stacks. Labelled images from $\mathcal{L}$ are fed into this stack, and its output is fed into a softmax layer whose other input is the class label. The complete scheme is depicted in Fig 3. Optimization is done through mini-batch stochastic gradient descent implemented through backpropagation with the Caffe package [12] (more details in Sec 4 and Supp).

4. Experiments

We validate our approach on 3 public datasets and compare to two existing methods, on equivariance (Sec 4.2), recognition performance (Sec 4.3) and next-best view selection (Sec 4.4). Throughout we compare the following methods:

- CLSNET: A neural network trained only from the supervised samples with a softmax loss.
- TEMPORAL: The *temporal coherence* approach of [21], which regularizes the classification loss with Eq (7) setting the distance measure $d(.)$ to the ℓ_1 distance in d_1. This method aims to learn invariant features by exploiting the fact that adjacent video frames should not change too much.
- DRLIM: The approach of [9], which also regularizes the classification loss with Eq (7), but setting $d(.)$ to the ℓ_2 distance in d_1.
- EQUIV: Our ego-motion equivariant feature learning approach, combined with the classification loss as in Eq (8), unless otherwise noted below.
- EQUIV+DRLIM: Our approach augmented with temporal coherence regularization ([9]).

TEMPORAL and DRLIM are the most pertinent baselines because they, like us, use contrastive loss-based formulations, but represent the popular "slowness"-based family of techniques ([36, 3, 8, 18]) for unsupervised feature learning from video, which, unlike our approach, are passive.

4.1. Experimental setup details

Recall that in the fully unsupervised mode, our method trains with pairs of video frames annotated only by their ego-poses in $\mathcal{U}$. In the supervised mode, when applied to recognition, our method additionally has access to a set of class-labeled images in $\mathcal{L}$. Similarly, the baselines all receive a pool of unsupervised data and supervised data. We now detail the data composing these two sets.

Unsupervised datasets We consider two unsupervised datasets, NORB and KITTI:
(1) **NORB** [15]: This dataset has 24,300 96×96-pixel images of 25 toys captured by systematically varying camera pose. We generate a random 67%-33% train-validation split and use 2D ego-pose vectors $\boldsymbol{y}$ consisting of camera elevation and azimuth. Because this dataset has discrete ego-pose variations, we consider two ego-motion patterns, *i.e.*, $G = 2$ (cf. Sec 3.1): one step along elevation and one step along azimuth. For EQUIV, we use all available positive pairs for each of the two motion patterns from the training images, yielding a $N_u = 45,417$-pair training set. For DRLIM and TEMPORAL, we create a 50,000-pair training set (positives to negatives ratio 1:3). Pairs within one step (elevation and/or azimuth) are treated as "temporal neighbors", as in the turntable results of [9, 21].
(2) **KITTI** [6, 7]: This dataset contains videos with registered GPS/IMU sensor streams captured on a car driving around 4 types of areas (location classes): "campus", "city", "residential", "road". We generate a random 67%-33% train-validation split and use 2D ego-pose vectors consisting of "yaw" and "forward position" (integral over "forward velocity" sensor outputs) from the sensors. We discover ego-motion patterns p_{ij} (cf. Sec 3.1) on frame pairs ≤ 1 second apart. We compute 6 clusters and automatically retain the $G = 3$ with the largest motions, which upon inspection correspond to "forward motion/zoom", "right turn", and "left turn" (see Fig 1, left). For EQUIV, we create a $N_u = 47,984$-pair training set with 11,996 positives. For DRLIM and TEMPORAL, we create a 98,460-pair training set with 24,615 "temporal neighbor" positives sampled $\leq$2 seconds apart. We use grayscale "camera 0" frames (see [7]), downsampled to 32×32 pixels, so that we can adopt CNN architecture choices known to be effective for tiny images [1].

Supervised datasets In our recognition experiments, we consider 3 supervised datasets $\mathcal{L}$: (1) **NORB**: We select 6 images from each of the $C = 25$ object training splits at random to create instance recognition training data. (2) **KITTI**: We select 4 images from each of the $C = 4$ location class training splits at random to create location recognition

Tasks→	Equivariance error		Recognition accuracy %				Next-best view
Datasets→	NORB		NORB-NORB	KITTI-KITTI	KITTI-SUN	KITTI-SUN	NORB
Methods↓	atomic	composite	[25 cls]	[4 cls]	[397 cls]	[397 cls, top-10]	1-view→ 2-view
random	1.0000	1.0000	4.00	25.00	0.25	2.52	4.00 → 4.00
CLSNET	0.9239	0.9145	25.11±0.72	41.81±0.38	0.70±0.12	6.10±0.67	-
TEMPORAL [21]	0.7587	0.8119	35.47±0.51	45.12±1.21	1.21±0.14	8.24±0.25	29.60→ 31.90
DRLIM [9]	0.6404	0.7263	36.60±0.41	47.04±0.50	1.02±0.12	6.78±0.32	14.89→ 17.95
EQUIV	**0.6082**	**0.6982**	**38.48±0.89**	**50.64±0.88**	**1.31±0.07**	**8.59±0.16**	**38.52→43.86**
EQUIV+DRLIM	**0.5814**	**0.6492**	**40.78±0.60**	**50.84±0.43**	**1.58±0.17**	**9.57±0.32**	**38.46→43.18**

Table 1. (Left) Average equivariance error (Eq (9)) on NORB for ego-motions like those in the training set (atomic) and novel ego-motions (composite). (Center) Recognition result for 3 datasets (mean ± standard error) of accuracy % over 5 repetitions. (Right) Next-best view selection accuracy %. Our method EQUIV (and augmented with slowness in EQUIV+DRLIM) clearly outperforms all baselines.

training data.(3) **SUN** [33]: We select 6 images for each of $C = 397$ scene categories at random to create scene recognition training data. We preprocess them identically to the KITTI images above (grayscale, crop to KITTI aspect ratio, resize to 32×32). We keep all the supervised datasets small, since unsupervised feature learning should be most beneficial when labeled data is scarce. Note that while the video frames of the unsupervised datasets U are associated with ego-poses, the static images of $\mathcal{L}$ have no such auxiliary data.

Network architectures and optimization For KITTI, we closely follow the cuda-convnet [1] recommended CIFAR-10 architecture: 32 conv(5x5)-max(3x3)-ReLU → 32 conv(5x5)-ReLU-avg(3x3) → 64 conv(5x5)-ReLU-avg(3x3) → D =64 full feature units. For NORB, we use a fully connected architecture: 20 full-ReLU→ D =100 full feature units. Parentheses indicate sizes of convolution or pooling kernels, and pooling layers have stride length 2.

We use Nesterov-accelerated stochastic gradient descent. The base learning rate and regularization λs are selected with greedy cross-validation. The contrastive loss margin parameter δ in Eq (6) is set to 1.0. We report all results for all methods based on 5 repetitions. For more details on architectures and optimization, see Supp.

4.2. Equivariance measurement

First, we test the learned features for equivariance. Equivariance is measured separately for each ego-motion g through the normalized error ρ_g:

$$\rho_g = E\left[\|\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) - M_g^{'}\mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})\|_2 / \|\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})\|_2\right], \tag{9}$$

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 [16] to solve for the equivariance maps of features produced by each compared method on held-out validation data, before computing ρ_g (see Supp).

We test both (1) "atomic" ego-motions matching those provided in the training pairs (*i.e.*, "up" 5°and "down" 20°) and (2) composite ego-motions ("up+right", "up+left", "down+right"). The latter lets us verify that our method's equivariance extends beyond those motion patterns used for training (cf. Sec 3.2). First, as a sanity check, we quantify equivariance for the unsupervised loss of Eq (5) in isolation, *i.e.*, learning with only $\mathcal{U}$. Our EQUIV method's average ρ_g error is 0.0304 and 0.0394 for atomic and composite ego-motions in NORB, respectively. In comparison, DRLIM—which promotes invariance, not equivariance—achieves $\rho_g = 0.3751$ and 0.4532. Thus, without class supervision, EQUIV tends to learn nearly completely equivariant features, even for novel composite transformations.

Next we evaluate equivariance for all methods using features optimized for the NORB recognition task. Table 1 (left) shows the results. As expected, we find that the features learned with EQUIV regularization are again easily the most equivariant. We also see that for all methods error is lower for atomic motions than composite motions, since they are more equivariant for smaller motions (see Supp).

4.3. Recognition results

Next we test the unsupervised-to-supervised transfer pipeline of Sec 3.4 on 3 recognition tasks: NORB-NORB, KITTI-KITTI, and KITTI-SUN. The first dataset in each pairing is unsupervised, and the second is supervised.

Table 1 (center) shows the results. On all 3 datasets, our method significantly improves classification accuracy, not just over the no-prior CLSNET baseline, but also over the closest previous unsupervised feature learning methods.[4]

All the unsupervised feature learning methods yield large gains over CLSNET on all three tasks. However, DRLIM and TEMPORAL are significantly weaker than the proposed method. Those methods are based on the "slow feature analysis" principle [31]—nearby frames must be close to one another in the learned feature space. We observe in practice (see Supp) that temporally close frames are mapped close to each other after only a few training epochs. This points to a possible weakness in these methods—even

[4]To verify the CLSNET baseline is legitimate, we also ran a Tiny Image nearest neighbor baseline on SUN as in [33]. It obtains 0.61% accuracy (worse than CLSNET, which obtains 0.70%).

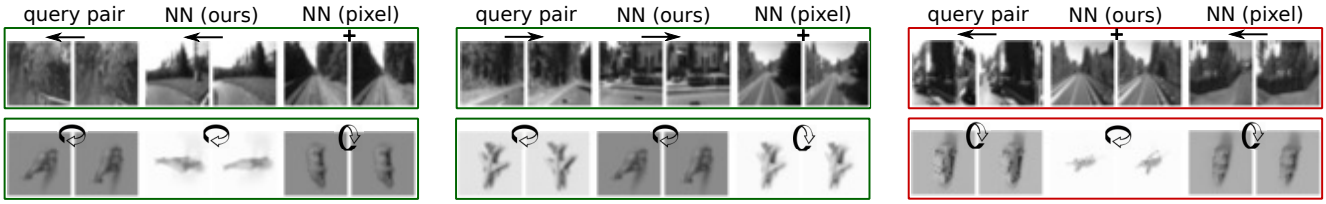

Figure 4. Nearest neighbor image pairs (cols 3 and 4 in each block) in pairwise equivariant feature difference space for various query image pairs (cols 1 and 2 per block). For comparison, cols 5 and 6 show pixel-wise difference-based neighbor pairs. The direction of ego-motion in query and neighbor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

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. EQUIV+DRLIM further improves over EQUIV, possibly because: (1) our EQUIV implementation only exploits frame pairs arising from specific motion patterns as positives, while DRLIM more broadly exploits all neighbor pairs, and (2) DRLIM and EQUIV losses are compatible— DRLIM requires that small perturbations affect features in small ways, and EQUIV requires that they affect them systematically.

The most exciting result is KITTI-SUN. The KITTI data itself is vastly more challenging than NORB due to its noisy ego-poses from inertial sensors, dynamic scenes with moving traffic, depth variations, occlusions, and objects that enter and exit the scene. Furthermore, the fact we can transfer EQUIV features learned without class labels on KITTI (street scenes from Karlsruhe, road-facing camera with fixed pitch and field of view) to be useful for a supervised task on the very different domain of SUN ("in the wild" web images from 397 categories mostly unrelated to streets) indicates the generality of our approach. Our best recognition accuracy of 1.58% on SUN is achieved with only 6 labeled examples per class. It is $\approx$30% better than the nearest competing baseline TEMPORAL and over 6 times better than chance. Top-10 accuracy trends are similar.

While we have thus far kept supervised training sets small to simulate categorization problems in the "long tail" where training samples are scarce and priors are most useful, new preliminary tests with larger labeled training sets on SUN show that our advantage is preserved. With N=20 samples for each of 397 classes on KITTI-SUN, EQUIV scored 3.66+/-0.08% accuracy vs. 1.66+/-0.18 for CLSNET.

4.4. Next-best view selection for recognition

Next, we show preliminary results of a direct application of equivariant features to "next-best view selection". Given one view of a NORB object, 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. Our method for this task, similar in spirit to [32], is described in detail in Supp. Table 1 (right) shows the results. On this task too, EQUIV features easily outperform the baselines.

4.5. Qualitative analysis

To qualitatively evaluate the impact of equivariant feature learning, we pose a 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). Fig 4 shows examples. For a variety of query pairs, we show the top neighbor pairs in the EQUIV space, as well as in pixel-difference space for comparison. Overall they visually confirm the desired equivariance property: neighbor-pairs in EQUIV's difference space exhibit a similar transformation (turning, zooming, *etc*.), whereas those in the original image space often do not. Consider the first azimuthal rotation NORB query in row 2, where pixel distance, perhaps dominated by the lighting, identifies a wrong ego-motion match, whereas our approach finds a correct match, despite the changed object identity, starting azimuth, lighting *etc*. The red boxes show failure cases. For instance, in the KITTI failure case shown (row 1, column 3), large foreground motion of a truck in the query image causes our method to wrongly miss the rotational motion.

5. Conclusion

Over the last decade, visual recognition methods have focused almost exclusively on learning from "bags of images". We argue that such "disembodied" image collections, though clearly valuable when collected at scale, deprive feature learning methods from the informative physical context of the original visual experience. We presented the first "embodied" approach to feature learning that generates features equivariant to ego-motion. Our results on multiple datasets and on multiple tasks show that our approach successfully learns equivariant features, which are beneficial for many downstream tasks and hold great promise for novel future applications.

Acknowledgements: This research is supported in part by ONR PECASE Award N00014-15-1-2291 and a gift from Intel.

References

[1] Cuda-convnet. https://code.google.com/p/cuda-convnet/. 6, 7

[2] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. Signature verification using a Siamese time delay neural network. *IJPRAI*, 1993. 5

[3] C. F. Cadieu and B. A. Olshausen. Learning intermediate-level representations of form and motion from natural movies. *Neural computation*, 2012. 3, 6

[4] C. Chen and K. Grauman. Watching unlabeled videos helps learn new human actions from very few labeled snapshots. In *CVPR*, 2013. 3

[5] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. *NIPS*, 2014. 2

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets Robotics: The KITTI Dataset. *IJRR*, 2013. 2, 4, 6

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. *CVPR*, 2012. 2, 6

[8] R. Goroshin, J. Bruna, J. Tompson, D. Eigen, and Y. LeCun. Unsupervised Learning of Spatiotemporally Coherent Metrics. *arXiv*, 2014. 2, 3, 6

[9] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. *CVPR*, 2006. 3, 5, 6, 7

[10] R. Held and A. Hein. Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology*, 1963. 1

[11] G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming Auto-Encoders. *ICANN*, 2011. 2, 3

[12] 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. 6

[13] J. J. Kivinen and C. K. Williams. Transformation equivariant boltzmann machines. *ICANN*, 2011. 2, 3, 4

[14] T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network. *arXiv*, 2015. 2, 3

[15] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. *CVPR*, 2004. 6

[16] K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. *CVPR*, 2015. 2, 3, 4, 7

[17] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 3

[18] J.-P. Lies, R. M. Häfner, and M. Bethge. Slowness and sparseness have diverging effects on complex cell learning. *PLoS computational biology*, 2014. 3, 6

[19] R. Memisevic. Learning to relate images. *PAMI*, 2013. 3

[20] V. Michalski, R. Memisevic, and K. Konda. Modeling Deep Temporal Dependencies with Recurrent Grammar Cells"". *NIPS*, 2014. 2, 3

[21] H. Mobahi, R. Collobert, and J. Weston. Deep Learning from Temporal Coherence in Video. *ICML*, 2009. 2, 3, 5, 6, 7

[22] T. Nakamura and M. Asada. Motion sketch: Acquisition of visual motion guided behaviors. *IJCAI*, 1995. 3

[23] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv*, 2014. 2, 3

[24] X. Ren and C. Gu. Figure-Ground Segmentation Improves Handled Object Recognition in Egocentric Video. In *CVPR*, 2010. 3

[25] U. Schmidt and S. Roth. Learning rotation-aware features: From invariant priors to equivariant descriptors. *CVPR*, 2012. 2, 3, 4

[26] P. Simard, Y. LeCun, J. Denker, and B. Victorri. Transformation Invariance in Pattern Recognition - Tangent distance and Tangent propagation. 1998. 2

[27] P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. *ICDAR*, 2003. 2

[28] K. Sohn and H. Lee. Learning invariant representations with local transformations. *ICML*, 2012. 2

[29] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and trends in computer graphics and vision*, 3(3):177–280, 2008. 2, 3

[30] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. *ICML*, 2008. 2

[31] L. Wiskott and T. J. Sejnowski. Slow feature analysis: unsupervised learning of invariances. *Neural computation*, 2002. 3, 7

[32] Z. Wu, S. Song, A. Khosla, X. Tang, and J. Xiao. 3d shapenets for 2.5 d object recognition and next-best-view prediction. *CVPR*, 2015. 8

[33] 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, 7

[34] C. Xu, J. Liu, and B. Kuipers. Moving object segmentation using motor signals. *ECCV*, 2012. 3

[35] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki. Attention prediction in egocentric video using motion and visual saliency. *PSIVT*, 2012. 3

[36] W. Zou, S. Zhu, K. Yu, and A. Y. Ng. Deep learning of invariant features via simulated fixations in video. *NIPS*, 2012. 3, 6