# *Comparing state-of-the-art visual features on invariant object recognition tasks*

# *Comparing State-of-the-Art Visual Features on Invariant Object Recognition Tasks*

Nicolas Pinto[1], Youssef Barhomi[1], David D. Cox[2], and James J. DiCarlo[1]

[1]*Massachusetts Institute of Technology, Cambridge, MA, U.S.A*
[2]*The Rowland Institute at Harvard, Cambridge, MA, U.S.A*

## Abstract

*Tolerance ("invariance") to identity-preserving image variation (e.g. variation in position, scale, pose, illumination) is a fundamental problem that any visual object recognition system, biological or engineered, must solve. While standard natural image database benchmarks are useful for guiding progress in computer vision, they can fail to probe the ability of a recognition system to solve the invariance problem [23, 24, 25]. Thus, to understand which computational approaches are making progress on solving the invariance problem, we compared and contrasted a variety of state-of-the-art visual representations using synthetic recognition tasks designed to systematically probe invariance. We successfully re-implemented a variety of state-of-the-art visual representations and confirmed their published performance on a natural image benchmark. We here report that most of these representations perform poorly on invariant recognition, but that one representation [21] shows significant performance gains over two baseline representations. We also show how this approach can more deeply illuminate the strengths and weaknesses of different visual representations and thus guide progress on invariant object recognition.*

## 1. Introduction

Visual object recognition is an extremely difficult problem and a great deal of effort continues to be expended to reach the goal of discovering visual representations that solve that problem (identification and categorization). Indeed, some of those representations are yielding performance that appears to be quite impressive [10, 22, 9, 34], perhaps even approaching human object recognition performance under limited conditions [30]. However, understanding what ideas are key to that progress, requires a clear focus on the computational crux problem and a critical, systematic evaluation of how much progress is being made on that problem by each state-of-the-art approach. The goal of the present study is to tackle this issue. For example, are current state-of-the-art representations all performing equally well, or are some consistently better than others? Do they each have weaknesses that might be overcome from learning from the strengths of each? Should we be satisfied with the single performance figure provided by a given natural images database, or can we more precisely determine what components of the object recognition problem are easily handled by each representation and what components are limiting performance?

The computational crux of object recognition is known as the "invariance" problem [33, 23]: any given object in the world can cast an essentially infinite number of different two-dimensional images onto the retina as the object's position, pose, lighting and background vary relative to the viewer. Thus, to critically evaluate a visual representation for object recognition, we must have ways of measuring its ability to solve the invariance problem. Even though performance evaluation is central in computer vision [3, 27], we do not believe that any previously study has directly and systematically tested state-of-the-art algorithms on solving the invariance problem.

In particular, some groups [20, 11] have employed tests that try to directly engage the invariance problem, but these test sometimes miss important components (e.g. failure to use appropriate backgrounds), and they have not been applied to compare and contrast state-of-the-art representations. Other groups [18, 19] have compared various visual descriptors (including SIFT, steerable filters, spin images or shape context), but the focus of these studies was not directly on the invariant object recognition problem, but on correspondence matching using local features similar in nature (i.e. "distribution-based" representations). A number of other recent evaluation studies (e.g. [34]), including a comprehensive study by [37] have evaluated the performance of state-of-the-art visual representations (and combinations of those representations) using so-called "natural" image databases (esp. Caltech-101 [8] or PASCAL VOC [7]). However, because image variation is not explicitly

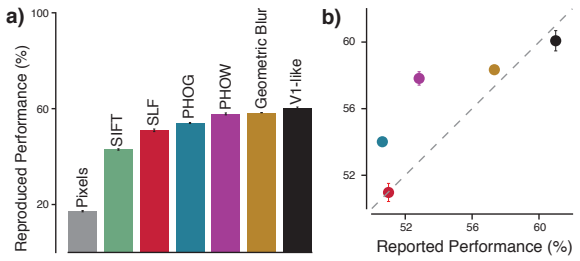


Figure 1. **Reproducing the state-of-the-art on Caltech-101.** a) Average accuracy with 15 training and 15 testing examples for five state-of-the-art algorithms and two baselines (see Methods). b) Reported vs reproduced performance showing the successful re-implementation of published methods. The reported numbers come from [21] (*SLF*), [35] (*PHOG*, *PHOW* and *Geometric Blur*) and [23] (*V1-like*).

controlled, these tests may lack real-world image variation (e.g. due to posing of photographs), making difficult or impossible to know how well the visual representations have solved the invariance problem. Moreover, performance on such tests may reflect successful exploitation of low-level regularities (e.g. due to covariation of object identity with background texture or color) and artifacts hidden in the test sets (e.g. cropping cues, etc.). While these problems have been pointed out in recent studies on natural image sets in object and face recognition [27, 23, 32, 24, 25], systematic tests that expose or circumvent them have not yet emerged.

To illuminate the progress of state-of-the-art visual representations in solving invariant object recognition, we re-implemented five state-of-the-art visual representations (some bio-inspired, some not), and we probed the ability of each representation to solve invariant object recognition tasks in which ground truth is known. Specifically, we used a synthetic image approach outlined in Figure 2 because it allows: parametric control of the invariance problem, control of shape similarity, control of the number of object exemplars in each category, control of background and color covariance. We compared the obtained performance with both a *Pixels* baseline representation and a well-established baseline representation that approximates the first level of primate visual processing (*V1-like* representation). We also used this approach to ask how well each visual representation handle each of these underlying types of invariance, which is difficult or impossible using prevailing "natural" object recognition tests.

# 2. Methods

## 2.1. Visual Representations

In the following, we give an overview of the visual features used in our experiments along with their key parameters. We refer to the corresponding publications for more details. Note that the terms "descriptors", "features" and "representations" are used interchangeably throughout the paper.

### 2.1.1 "Baseline" Features

We used two simple image representations – *Pixels* and *V1-like* – designed to serve as baselines against which the performance of state-of-the-art features can be measured. For both baseline representations, training and testing data were normalized to have zero-mean and unit-variance feature-wise (using the training data only), and a simple linear kernel was used for classification (see Section 2.1.3).

**Pixels:** In the *Pixels* representation, each image was simply rescaled to 150 by 150 pixels, converted to grayscale and then unrolled as a feature vector. The resulting feature vector represents an almost entirely unprocessed representation of the original image.

**V1-like:** In the *V1-like* representation, features were taken without any additional optimization from Pinto et al.'s V1S+ [23]. This visual representation consists of a collection of locally-normalized, thresholded Gabor wavelet functions spanning a range of orientations and spatial frequencies and is based on a first-order description of primary visual cortex V1. *V1-like* features have been proposed by neuroscientists as a "null" model for object recognition since they do not contain a particularly sophisticated representation of shape or appearance, nor do they possess any explicit mechanism designed to tolerate image variation (e.g. changes in view, lighting, position, etc. [5, 23]). In spite of their simplicity, these features have been shown to be among the best-performing non-blended features set on standard natural face and object recognition benchmarks (i.e. Caltech-101, Caltech256, ORL, Yale, CVL, AR, PIE, LFW [23, 24, 25]), and are a key component of the best blended solutions for some of these same benchmarks [22]. We used publicly available code for these features with two minor modifications to the published procedure. Specifically, no PCA dimensionality reduction was performed prior to classification (the full vector was used) and a different regularization parameter was used ($C = 10,000$ instead of $C = 10$).

### 2.1.2 State-of-the-art Features

We considered a diverse set of five state-of-the-art features. Most were chosen on the basis of their high-performance on Caltech-101 (arguably still the most widely used multi-class object recognition benchmark today [10, 9, 22]). Effort was made to span a wide range of different approaches to object recognition: models that were bio-inspired, and those that

are not; distribution-based and non-distribution-based models, and models with a custom kernel (e.g. Spatial Pyramid) and models with a simple linear one. To promote experimental reproducibility and ease distribution, we re-implemented all but one of these models (*SLF*, see below) from the ground up using only free, open-source software (e.g. Python, NumPy, SciPy, Shogun, OpenCV, etc.).

**SIFT:** *SIFT* descriptors [15] were computed on a uniform dense grid from a 150 by 150 pixels grayscale image with a spacing of 10 pixels and a single patch size of 32 by 32 pixels. The result was then unwrapped as a feature vector. Training and testing data were normalized to have zero-mean and unit-variance feature-wise (using the training data only), and SVM classification was done using a linear kernel.

**PHOW:** *PHOW* (Pyramid Histogram Of visual Words) is a spatial pyramid representation of appearance [2, 35, 12]. To compute these features, a dictionary of visual words was first generated by quantizing the *SIFT* descriptors with k-means clustering. We fixed the dictionary size to 300 elements, and the SVM kernel to a three-level spatial pyramid kernel with $\chi^2$ distance [12].

**PHOG:** *PHOG* (Pyramid Histogram Of Gradients) is a spatial pyramid representation of shape [2, 35] based on orientations gradients (HOG [4]) of edges extracted with a Canny detector. We fixed the angular range to 360 degrees, the number of quantization bins to 40, and the SVM kernel to a four-level spatial pyramid kernel with $\chi^2$ distance.

**Geometric Blur:** The *Geometric Blur* shape descriptors [1, 36] are generated by applying spatially varying blur on the surrounding patch of edge points in the image (extracted by the boundary detector of [16]). We fixed the blur parameters to $\alpha = 0.5$ and $\beta = 1$, the number of descriptors to 300 and the maximum radius to 50 pixels. For the SVM classification, we used the kernelized distance $D^A$ from [36] (Eq. 1) with no texture term as described in [35].

**SLF:** The bio-inspired *Sparse Localized Features* (SLF) [21] are an extensions of the *C2* features from the Serre et al. HMAX model [31, 28]. For this representation, we took advantage of the MATLAB code provided by the authors. Here, the SVM classification was based on a linear kernel with normalized training and testing data (zero-mean and unit-variance feature-wise). Interestingly, we found that it was unnecessary to use the feature selection procedure described in [21] to match the level of Caltech-101 performance achieved in that work. We suspect that our slightly higher observed performance level was due to differences in SVM formulation and regularization parameters.

#### 2.1.3 Classification

For classification we used L2-regularized Support Vector Machines (libsvm solver from the Shogun Toolbox[1]) with a regularization constant $C = 10,000$. Each representation was used to produce either a simple linear, or custom (for *PHOW*, *PHOG* and *Geometric Blur*) kernel. Multi-class problems were addressed with a one-versus-rest formulation.

Classifiers were trained using a fixed number of examples. Except when stated otherwise, we use 150 training and 150 testing examples for each class. The performance scores reported are the average of performances obtained from five random splits of training and testing sets, the error bars represent the standard error of the mean. The same image splits were used for all the representations.

### 2.2. Synthetic Image Set Generation

A key feature of the evaluation procedure described in this study is the use of object test sets where the ground-truth range of variation in object view is known. In particular, we chose to use rendered three-dimensional objects, which allow for large numbers of test images to be generated with minimal effort, while preserving tight controls on the distribution of view variation within the set.

For each category of objects (cars, planes, boats, animals), five 3D meshes (purchased from Dosch Design and Turbosquid.com) were rendered using the POV-Ray ray-tracing package onto a transparent background, and this image was overlaid onto a randomly selected background image from a set of more than 2,000 images of natural scenes (Figure 2a). Background images were selected randomly, and no background was ever reused within a given training / test set. While backgrounds often contain information that is helpful for recognizing objects, we made no effort to associate objects with context-appropriate background, in order to better focus the test set on object recognition *per se*. All images were made to be grayscale to avoid any color confound.

In Figures 3a and 4a, object views were varied simultaneously ("composite variation") along four axes: position (horizontal and vertical), scale, in-plane rotation and in-depth rotation. In order to roughly equate the effects of each of these kinds of view variation, we defined a view change "quantum" for each axis of variation, such that each kind of variation, on average, produced an equivalent pixel change in the image, as defined by a pixel-wise Euclidean distance. The average pixel change associated with a full, non-overlapping translations of the objects' bounding boxes

[1] http://www.shogun-toolbox.org

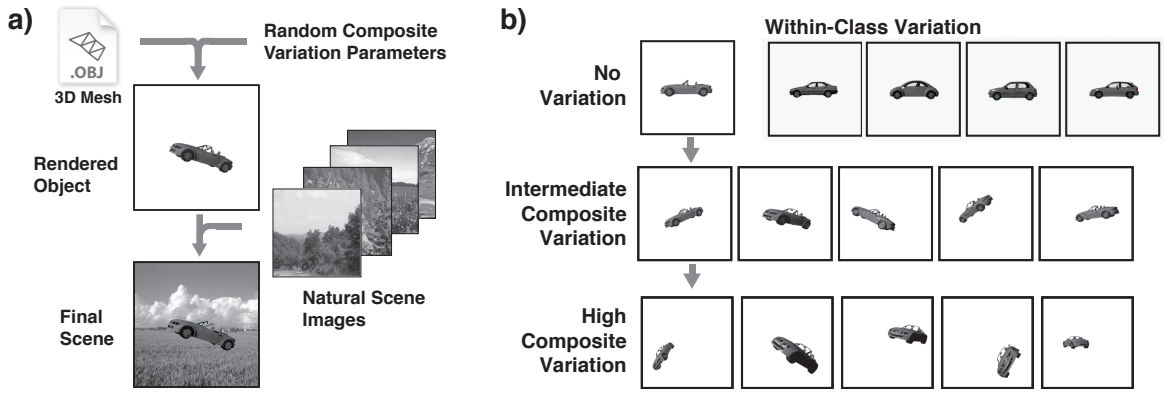


Figure 2. **Object rendering procedure.** a) 3D object meshes were rendered using view parameters drawn from a uniform random distribution, and then composited onto randomly selected natural background images. b) Examples of view variation ranges used in this study, spanning from no view variation (top) to relatively large amounts of "composite variation" (i.e. all four types of variation included: position, scale, in-plane rotation and in-depth rotation).

was taken as the "standard" unit of pixel variation, and all other view change units were equalized to this unit.

Separate test and training sets were generated for each of a series of view-variation ranges, spanning from no view variation (Figure 2b, top) to a relatively large amount of variation (Figure 2b, bottom). For each range, view parameters were drawn independently along each of the four axes, with a uniform random distribution, and all object exemplars were included as part of the random image draw. Importantly, for a given range, a successful recognition system must not only correctly recognize objects with view parameters at the extremes of the range, but must also correctly recognize objects across the entire range.

## 3. Results

The main goal of this study was to test state-of-the-art artificial visual representations on truly difficult, systematic tests of invariant object recognition where ground truth is known [23]. To do this, we first verified that we had successfully re-implemented each visual representations (see Methods). We used the Caltech-101 image categorization task [8] as a point of reference. Despite many serious concerns raised about the Caltech-101 set [27, 23], that test is still widely used in the object recognition community and thus most state-of-art algorithms have reported accuracy on Caltech-101 in the literature [1, 36, 35, 21, 12]. Specifically, for each representation, we compared the Caltech-101 performance of our re-implementation with the performance reported in the literature. As Figure 1b shows, in all cases, we succeeded in matching (or slightly exceeding) the reported performance of all representations on the Caltech-101 set. These results provide an independent replication of the original authors' results, and that we have succeeded in re-implementing these state-of-the-art algorithms.

### 3.1. Basic-level Object Recognition

With successful re-implementations of a collection of state-of-the-art algorithms in hand, we proceeded to test each representation on basic-level invariant object recognition tasks and to benchmark these results against a simple *Pixels* baseline representation and the *V1-like* baseline representation (see *Methods*).

In Figure 3, we show the performance of all seven visual representations (including the *Pixels* and *V1-like* baseline representations) as we gradually increase the difficulty of the "cars vs. planes" task by increasing four types of object variation (position, scale, in-plane rotation and in-depth rotation) at the same time in a fixed mixture ("composite invariance", see Methods Section 2.2). Even though all of the state-of-the-art representations consistently outperformed *Pixels* and performed approximately equally well on the standard Caltech-101 "natural" task (Figure 1), the results in Figure 3 reveal clear differences among the models. Several state-of-the-art models are clearly below the *V1-like* baseline and, in some case, below the *Pixels* baseline. Most interestingly, the results show that one representation (*SLF*) has made clear gains on the composite invariance problem (see *Discussion*).

Given that there is no single test of basic-level recognition, we next considered the possibility that the results in Figure 3a are simply due to particular parameter choices one necessarily has to make when testing a visual representation (e.g. number of training examples, number of objects, particular choice of objects, etc). Specifically, we picked an intermediate level of composite variation that best revealed the differences among the representations (see highlighted section in Figure 3a and, using this level of compos-

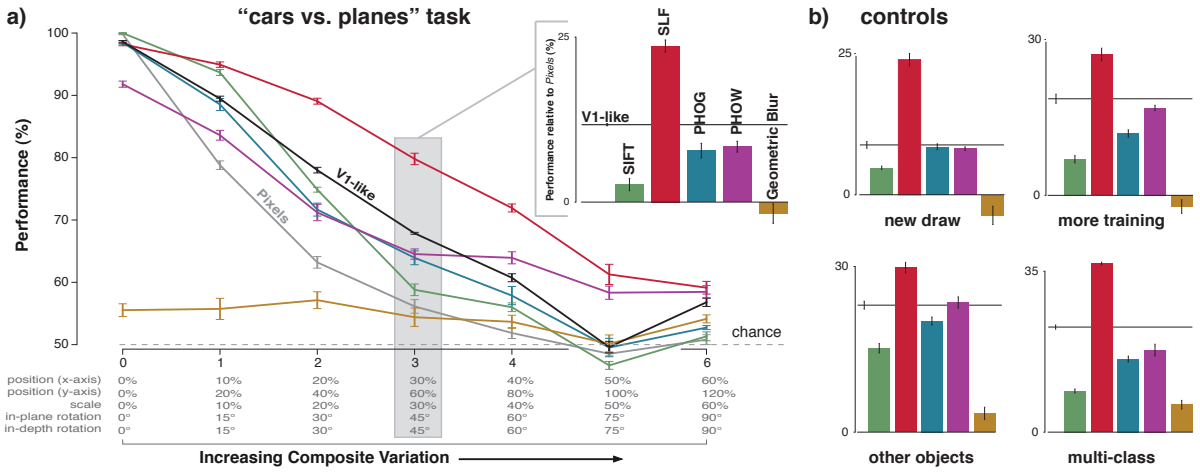


Figure 3. **Performance on object recognition tasks with controlled composite variation.** a) Average accuracy of each representation on a series of "cars vs planes" tasks in which the composite variation is gradually increased. The inset shows the performance of the five state-of-the-art features from the literature relative to the *Pixels* representation. b) Performance relative to *Pixels* on the composite variation 3 (cf. inset in a)) with a new draw of "cars" and "planes" images, more training examples, other objects ("animals vs boats"), and more object classes ("animals vs boats vs cars vs planes").

ite variation, we created four new object recognition tests using: a new set of "cars" and "planes" images, three times as many example images for training (450 instead of 150), two completely new objects ("animals" and "boats"), and a test with four basic object categories ("animals", "boats", "cars", "planes") instead of two (Figure 3b). In all cases, we found that the relative performance of each representation (i.e. performance relative to the other representations) was largely unaffected by these testing parameter choices. We quantified this robustness by computing Spearman's rank correlation of performance in all pairs of these basic level recognition tasks, and found very high values (mean = 0.95, min = 0.86). In sum, these results show that, at least for the currently considered set of state-of-the-art models, our tests of basic-level recognition are largely robust to: the exact set of images (at a given level of composite variation) the number of training examples, the exact categories of basic object used and the number of categories.

### 3.2. Subordinate-level Object Recognition (Faces)

The absolute level of recognition performance must depend on the degree of 3D structural similarity of the objects in the test set. Specifically, while objects involved in tests of basic-level recognition (e.g. cars vs. planes vs. boats, etc.) are moderately to highly dissimilar in terms of 3D structure, objects in so-called, subordinate-level [17] tasks of recognition (e.g. one face vs. another face) share common 3D structure that makes tasks intrinsically more challenging. Thus, we used the same approach as in Figure 3 (but with lower absolute levels of view variance) to test the performance of the state-of-the-art representations on a face recognition task. The results are shown in Figure 4. As with the basic-level recognition task, we found that most, but not-all, state-of-the-art representations performed below the *V1-like* baseline representation and that the relative performance of the representations on the face task was largely robust to the number of training examples, the particular choice of faces, and the number of faces (mean Spearman's rank correlation = 0.92, min = 0.85).

To ask if a representation's performance on basic-level recognition is predictive of its performance on subordinate-level recognition, we directly compared the results in the "cars vs. planes" task (Figure 3) and the "face vs. face" task (Figure 4). Figure 5 shows the performance of all representations on both tasks using a range of different testing conditions (as outlined in Figures 3 and 4). We found that the absolute performance level on each task is highly correlated (Figure 5a). However, when performance is plotted relative to the *Pixels* (Figure 5b) and *V1-like* baselines (Figure 5c), the data reveal that one of the state-of-the-art representations (*PHOW*, see purple points) is reasonably good at basic-level invariant object recognition but quite poor at subordinate-level (face) recognition, and that one representation (*SLF*, see red points) is a clear stand-out with respect to the *V1-like* baseline on both basic- and subordinate-level recognition tasks (see *Discussion*).

### 3.3. Individual Types of View Variation

We next considered the possibility that our tests (e.g. Figure 3) were over-weighting some types of variation relative to others (see Methods Section 2.2, and [23] for a description of how the relative mix of types of object varia-

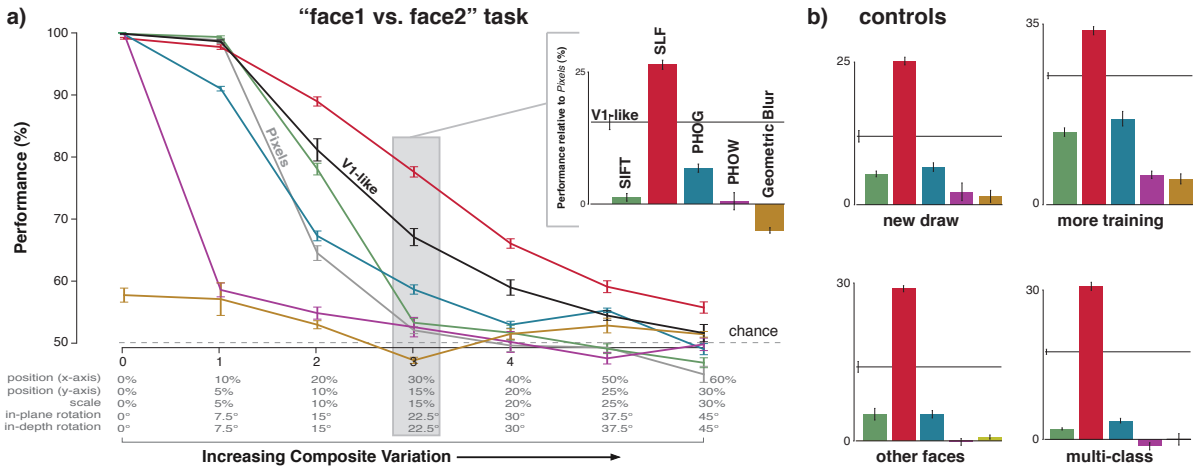


Figure 4. **Performance on subordinate-level object recognition tasks (face discrimination) with gradually increasing amounts of composite variation.** Plotting conventions as in Figure 3, but note that, because this is a more difficult that, view variation parameters are lower than those used in Figure 3.

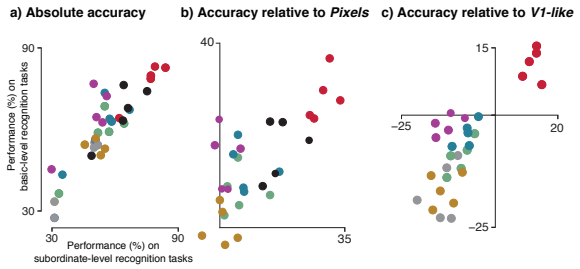


Figure 5. **Comparison of performance on basic-level and subordinate-level object recognition tasks.** a) We plot the accuracy of the representations on five tasks at composite variation 3 (original, new draw, more training, other objects and more objects; see Figures 3 and 4 for details). b) Same data re-plotted relative to *Pixels* representation. c) Same data re-plotted relative to *V1-like* representation.

tion in Figure 3 was chosen). Without an operational goal (e.g. matching human performance), it is impossible to exactly determine if one type of variation is under- or overweighted in recognition tasks, even when ground truth is known. However, for the present study, we sought to determine if our conclusions about the relative performance of the state-of-the-art representations would be strongly altered by our current weighting of each of these four type of view variation. Specifically, we created four new basic-level recognition tests in which we fully removed one type of variation from each test (and made sure that the remaining composite variation was at a level that put all the representations in a performance regime that was not on the ceiling or the floor, analogous to the highlighted region in Figure 3). We found that the relative performance of all the state-of-the-art representations on these four basic-level ob-

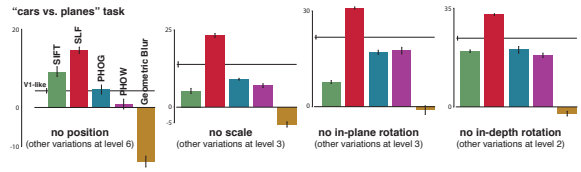


Figure 6. **Performance rank order of the representations after removal of each type of view variation (position, scale, in-plane rotation and in-depth rotation).**

ject recognition tasks (Figure 6) was very similar to that found with the full composite invariance tests (Figure 3; mean Spearman's rank correlation between results in these two figures was 0.90, min = 0.64). This suggests that, at least for the currently considered set of state-of-the-art models, our tests of basic-level recognition are not strongly dependent on composite variation "mixture" in the test.

To ask what type of tolerance is least difficult and most difficult for each representation without regard to absolute performance, we created four new tests of basic level recognition ("cars vs. planes") that each contained only one type of object variation (position-only, scale-only, in-plane-rotation-only, and in-depth-rotation-only tests). To fairly compare each representation's degree-of-difficulty in handling each type of variation, we equated these four tests in that the amount of variation produced an equal degree of difficulty for the pixel representation (10% absolute performance drop, see Figure 7). The results show that most of the state-of-the-art representations have the least difficulty with position variation, and tend to have more difficulty with (e.g.) in-depth variation. For example, these tests reveal that the representation of [21] which was designed with position and scale variation in mind [28, 31] is much less sensitive

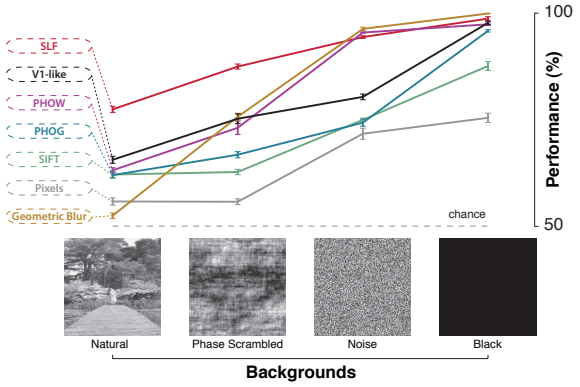


Figure 8. **Effect of the background type on each representation's performance on a basic level recognition task ("cars vs. planes") at composite object variation 3 (see Figure 3a).** More natural backgrounds (more "clutter") produce a higher degree of difficulty for all representations. Note that, although *Geometric Blur* and *PHOW* look very promising with simple backgrounds, they are particularly disrupted by natural backgrounds that are uncorrelated with object identity. .

to position variation than it is to in-depth rotation (i.e. pose) variation.

### 3.4. The influence of background

Because background structure and its covariance with object identity are fully known, the testing methods used here can also expose visual representations that rely strongly on these cues. For example, Figures 3 and 4 show that one of the state-of-the-art visual representations, *Geometric Blur*, performs very poorly on most of our tests, but Figure 8 shows that, when we perform the tests on no background, the same representation now performs at a very high level. Taken together, this suggests that this visual representation is seriously impaired by clutter or leans heavily on background features to perform categorization. When natural images are used and background covariance is brought to zero (as in all our testing), this limitation of the representation is revealed. We emphasize that these effects are difficult or impossible to uncover in standard "natural" tests, but are very easy to uncover using a synthetic test set approach (see *Discussion*).

## 4. Discussion

Our testing of invariant object recognition revealed that most of the state-of-the-art representations consistently performed at or below the performance of the *V1-like* baseline representation (which also achieves the highest performance on the Caltech-101 set). To the extent that this model represents a "null" baseline that lacks mechanisms to perform invariant recognition, this suggests that other state-of-the-art representations perhaps also rely heavily on view-specific information, or covariation with backgrounds, to achieve their performance. Interestingly, the bio-inspired model *SLF* (an extension of Serre et al.'s C2 features [31]) stood out in all of our tests, performing consistently better than both baselines, suggesting that it contains computational ideas that are useful for solving invariant object recognition. It remains to be seen how this visual representation and other emerging representations compare to unfettered human performance and the performance of high-level neuronal representations on these tasks.

Our results also revealed that the performance of some representations was highly dependent on the details of the task under test. For example, the performance of *Geometric Blur* descriptors degraded rapidly with the inclusion of background content uncorrelated to object identity, and *PHOW*, while reasonably good at basic-level object categorization tasks, was no better than the pixel representation at the subordinate-level task (face identification).

The synthetic testing approach used here is partly motivated by previous work on photographic approaches like NORB [13]. However, while NORB-like databases are challenging and costly to obtain, the synthetic approach offers the potential to draw on large numbers of objects and generate an essentially infinite number of images with precise control of all key variables at low cost. The approach easily allows exploration of the individual underlying difficulty variables (e.g. position, scale, background, etc.) to better learn from the best ideas of each representation. Because the synthetic approach offers the ability to gradually "ratchet up" the task difficulty (e.g. increasing levels of composite variation in the test) and because only hundreds of images are needed instead of thousands, it can be used to efficiently search for better visual representations [26].

Although it is widely understood that performance evaluation is critical to driving progress, such performance evaluation is much easier said than done. Over the last decade, tests based on known ground truth have fallen out of fashion in computer vision [6] while a number of "natural" image test sets (e.g. Caltech-101, PASCAL VOC) continue to be used almost exclusively as evidence of progress in solving object recognition [10, 9, 22, 34]. While these test sets are laudable because they encourage systematic comparison of various algorithms, they can also be dangerous when hidden confounds exist in the sets, or when it is not clear *why* the sets are difficult. Indeed, despite the fact that these representations are highly competitive on large, complex "natural" image sets and the expectation [14] that many of these representations should be capable of dealing "fairly well" with simpler synthetic invariance tests, our results show that many of these representations are surprisingly weak on these tests, even though these synthetic sets remain trivially easy for human observers.

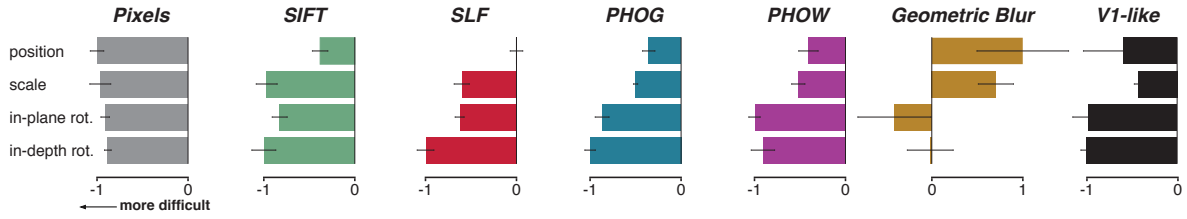


Figure 7. **Degree of difficulty of each type of view variation for each representation.** The amount of each type of variation was chosen to produce a 10% approximate decrease in performance for the *Pixels* baseline. We here show the change in average performance due to each type of variation (i.e. change relative to each representation's performance in the "no variation" task (composite variation 0) in Figure 3). The axes are normalized by the maximum decrease (or increase, for *Geometric Blur*) of performance for each representation. Thus, each plot shows the relative degree of difficulty for each type of variation (from the representation's point of view; -1 is most difficult). Even though this figure suggests that *Geometric Blur* benefits from more position and scale variations, that is only a by-product of the overall poor performance of this representation and floor effects (see Figure 3).

While there is no perfect evaluation tool, we believe that a synthetic testing approach is an important complement to ever-improving photographic-based image sets (e.g. LabelMe [29]). We share the desire and ultimate goal of evaluating algorithms on "real-world" tasks, and we share the concern that ill-considered synthetic testing approaches may not be predictive of real world performance. But in the world of modern computer graphics, a synthetic testing approach offers a powerful path forward as it can ultimately produce images that are indistinguishable from real-world photographs, yet still have all ground truth variables known and under parametric control.

## References


[1] A. C. Berg and J. Malik. Geometric blur and template matching. *CVPR*, 2001. 3, 4

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. *CIVR*, 2007. 3

[3] H. Christensen and P. Phillips. *Empirical evaluation methods in computer vision*. 2002. 1

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 3

[5] J. J. DiCarlo and D. D. Cox. Untangling invariant object recognition. *TICS*, 2007. 2

[6] S. Dickinson. The evolution of object categorization and the challenge of image abstraction. *Object Categorization: Computer and Human Vision Perspectives.*, 2009. 7

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. 1

[8] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *CVPR*, 2004. 1, 4

[9] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *ICCV*, 2009. 1, 2, 7

[10] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *CVPR*, 2009. 1, 2, 7

[11] S. Kim and I. Kweon. Biologically motivated perceptual feature: Generalized robust invariant feature. *ACCV*, 2006. 1

[12] S. Lazebnik, C. Schmid, and J. Ponce. Spatial Pyramid Matching. *Object Categorization: Computer and Human Vision Perspectives.*, 2009. 3, 4

[13] Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR*, 2004. 7

[14] Y. LeCun, D. G. Lowe, J. Malik, J. Mutch, P. Perona, and T. Poggio. Object recognition, computer vision, and the caltech 101: A response to pinto et al., 2008. 7

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 3

[16] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 2004. 3

[17] C. Mervis and E. Rosch. Categorization of natural objects. *Annual review of psychology*, 1981. 5

[18] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 2005. 1

[19] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *ICCV*, 2007. 1

[20] H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearance. *IJCV*, 1995. 1

[21] J. Mutch and D. G. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *IJCV*, 2008. 1, 2, 3, 4, 6

[22] Peter V. Gehler and Sebastian Nowozin. On Feature Combination for Multiclass Object Classification. In *ICCV*, 2009. 1, 2, 7

[23] N. Pinto, D. D. Cox, and J. J. DiCarlo. Why is Real-World Visual Object Recognition Hard. *PLoS Comp. Bio.*, 2008. 1, 2, 4, 5

[24] N. Pinto, J. J. DiCarlo, and D. D. Cox. Establishing Good Benchmarks and Baselines for Face Recognition. *ECCV*, 2008. 1, 2

[25] N. Pinto, J. J. DiCarlo, and D. D. Cox. How far can you get with a modern face recognition test set using only simple features? *CVPR*, 2009. 1, 2

[26] N. Pinto, D. Doukhan, J. J. DiCarlo, and D. D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comp Bio*, 2009. 7

[27] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, et al. Dataset Issues in Object Recognition. *LNCS*, 2006. 1, 2, 4

[28] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 1999. 3, 6

[29] B. Russell, A. Torralba, K. Murphy, and W. Freeman. LabelMe: a database and web-based tool for image annotation. *IJCV*, 2008. 8

[30] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *PNAS*, 2007. 1

[31] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *PAMI*, 2007. 3, 6, 7

[32] L. Shamir. Evaluation of Face Datasets as Tools for Assessing the Performance of Face Recognition Methods. *IJCV*, 2008. 2

[33] S. Ullman. *High-level vision: Object recognition and visual cognition*. The MIT Press, 2000. 1

[34] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 2010. 1, 7

[35] M. Varma and D. Ray. Learning The Discriminative Power-Invariance Trade-Off. *ICCV*, 2007. 2, 3, 4

[36] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006. 3, 4

[37] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007. 1