Neuron
Article

Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex

Nuo Li[1] and James J. DiCarlo[1,*]
[1]McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: dicarlo@mit.edu
DOI 10.1016/j.neuron.2010.08.029

SUMMARY

We easily recognize objects and faces across a myriad of retinal images produced by each object. One hypothesis is that this tolerance (a.k.a. "invariance") is learned by relying on the fact that object identities are temporally stable. While we previously found neuronal evidence supporting this idea at the top of the nonhuman primate ventral visual stream (inferior temporal cortex, or IT), we here test if this is a general tolerance learning mechanism. First, we found that the same type of unsupervised experience that reshaped IT position tolerance also predictably reshaped IT size tolerance, and the magnitude of reshaping was quantitatively similar. Second, this tolerance reshaping can be induced under naturally occurring dynamic visual experience, even without eye movements. Third, unsupervised temporal contiguous experience can build new neuronal tolerance. These results suggest that the ventral visual stream uses a general unsupervised tolerance learning algorithm to build its invariant object representation.

INTRODUCTION

Our ability to recognize objects and faces is remarkably tolerant to variation in the retinal images produced by each object. That is, we can easily recognize each object even though it can appear in different positions, sizes, poses, etc. In the primate brain, the solution to this "invariance" problem is thought to be achieved through a series of transformations along the ventral visual stream. At the highest stage of this stream, the inferior temporal cortex (IT), a tolerant object representation is obtained in which individual IT neurons have a preference for some objects ("selectivity") over others, and this rank-order preference is largely maintained across identity-preserving image transformations (Ito et al., 1995; Logothetis and Sheinberg, 1996; Tanaka, 1996; Vogels and Orban, 1996). Though most IT neurons are not strictly "invariant" (DiCarlo and Maunsell, 2003; Ito et al., 1995; Logothetis and Sheinberg, 1996; Vogels and Orban, 1996), reasonably sized populations of these so-called "tolerant" neurons can support object recognition tasks (Afraz et al., 2006; Hung et al., 2005; Li et al., 2009). However, we do not yet understand how IT neurons construct this tolerant response phenomenology.

One potentially powerful idea is that time can act as an implicit teacher, in that the temporal contiguity of object features during natural visual experience can instruct the learning of tolerance, potentially in an unsupervised manner (Foldiak, 1991; Masquelier et al., 2007; Masquelier and Thorpe, 2007; Sprekeler et al., 2007; Stryker, 1991; Wiskott and Sejnowski, 2002; Wyss et al., 2006). The overarching logic is as follows: during natural visual experience, objects tend to remain present for seconds or more, while object motion or viewer motion (e.g., eye movements) tend to cause rapid changes in the retinal image cast by each object over shorter time intervals (hundreds of ms). In theory, the ventral stream could construct a tolerant object representation by taking advantage of this natural tendency for temporally contiguous retinal images to belong to the same object, thus yielding tolerant object selectivity in IT cortex. A recent experimental result in adult nonhuman primate IT has provided some neuronal support for this temporal contiguity hypothesis (Li and DiCarlo, 2008). Specifically, we found that alterations of unsupervised experience of temporally contiguous object image changes across saccadic eye movements can induce rapid reshaping (within hours) of IT neuronal position tolerance (i.e., a reshaping of each IT neuron's ability to respond with consistent object selectivity across the retina). This IT neuronal learning likely has perceptual consequences because similar temporal contiguity manipulations of eye-movement-driven position experience can produce qualitatively similar changes in the position tolerance of human object perception (Cox et al., 2005).

However, these previous studies have two key limitations. First, they only uncovered evidence for temporal contiguity learning under a very restricted set of conditions: they showed learning effects only in the context of eye movements, and they only tested one type of tolerance—position tolerance. Because eye movements drive a great deal of the image statistics relevant only to position tolerance (temporally contiguous image translations), the previous results could reflect only a special case of tolerance learning. Second, the previous studies did not directly show that temporally contiguous image statistics can *build* new tolerance, but only showed that alterations of

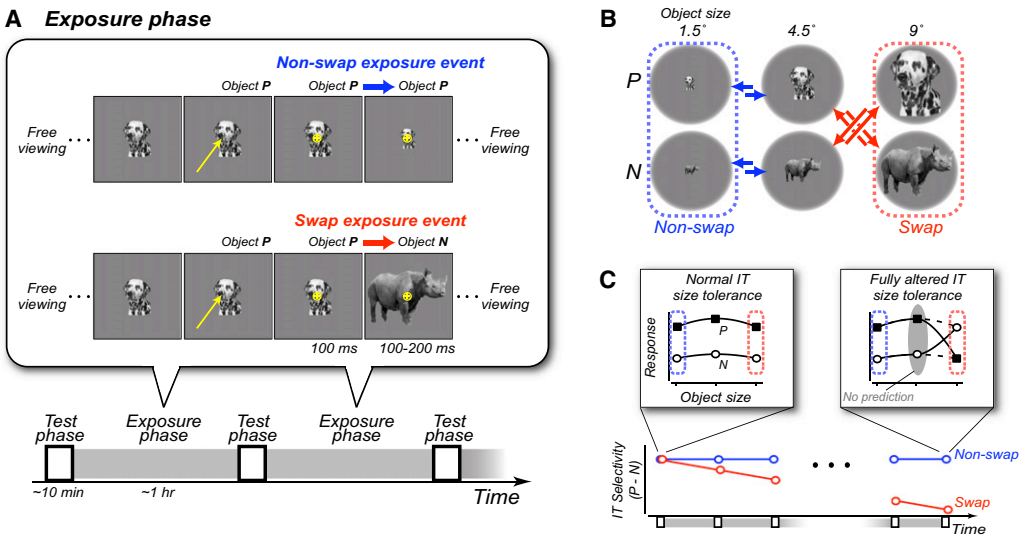


Figure 1. Experimental Design and Prediction

(A) IT selectivity was tested in the *Test Phases* whereas animals received experience in the altered visual world in the *Exposure Phases*.

(B) The chart shows the full exposure design for a single IT site in Experiment I. Arrows show the temporal contiguity experience of retinal images (arrow heads point to the retinal images occurring later in time, e.g., A). Each arrow shows a particular exposure event type (i.e., temporally linked images shown to the animal), and all eight exposure event types were shown equally often (randomly interleaved) in each *Exposure Phase*.

(C) Prediction for IT responses collected in the *Test Phase*: if the visual system builds size tolerance using temporal contiguity, the swap exposure should cause incorrect grouping of two different object images (P and N). The qualitative prediction is a decrease in object selectivity at the swap size (images and data points outlined in red) that grows stronger with increasing exposure (in the limit, reversing object preference as illustrated schematically here), and little or no change in object selectivity at the non-swap size. The experiment makes no quantitative prediction for the selectivity at the medium size (gray oval, see text).

those statistics can disrupt normal tolerance. Because of these limitations, we do not know if the naive ventral stream uses a general, temporal contiguity-driven learning mechanism to construct its tolerance to all types of image variation.

Here, we set out to test the temporal contiguity hypothesis in three ways. First, we reasoned that, if the ventral stream is using temporal contiguity to drive a general tolerance-building mechanism, alterations in that temporal contiguity should reshape other types of tolerance (e.g., size tolerance, pose tolerance, illumination tolerance), and the magnitude of that reshaping should be similar to that found for position tolerance. We decided to test size tolerance, because normal size tolerance in IT is much better described (Brincat and Connor, 2004; Ito et al., 1995; Logothetis and Sheinberg, 1996; Vogels and Orban, 1996) than pose or illumination tolerance. Our experimental logic follows our previous work on position tolerance (Cox et al., 2005; Li and DiCarlo, 2008). Specifically, when an adult animal with a mature (e.g., size-tolerant) object representation is exposed to an altered visual world in which object identity is consistently swapped across object size change, its visual system should learn from those image statistics such that it predictably "breaks" the size tolerance of that mature object representation. Assuming IT conveys this object representation (Afraz et al., 2006; Hung et al., 2005; Logothetis and Sheinberg, 1996; Tanaka, 1996), that learning should result in a specific change in the size tolerance of mature IT neurons (Figure 1).

Second, many types of identity-preserving image transformations in natural vision do not involve intervening eye movements (e.g., object motion producing a change in object image size). If the ventral stream is using a general tolerance-building mechanism, we should be able to find size tolerance reshaping even without intervening eye movements, and we should also be able to find size tolerance reshaping when the dynamics of the image statistics mimic naturally occurring image dynamics.

Third, our previous studies (Cox et al., 2005; Li and DiCarlo, 2008) and our first two aims above use the breaking of naturally occurring image statistics to try to break the normal tolerance observed in IT (i.e., to weaken existing IT object selectivity in a position- or size-specific manner; Figure 1). Such results support the inference that naturally occurring image statistics instruct the "building" of that tolerance in the naive ventral stream. However, we also sought to test that inference more directly by looking for evidence that temporally contiguous image statistics can build new tolerance in IT neurons with immature tolerance (i.e., can produce an increase in existing IT object selectivity in a position- or size-specific manner).

Our results showed that targeted alterations in the temporal contiguity of visual experience robustly and predictably reshaped IT neuronal size tolerance over a period of hours. This change in size tolerance grew gradually stronger with increasing visual experience, and the rate of reshaping was very similar to previously reported position tolerance reshaping (Li and DiCarlo, 2008). Second, we found that the size tolerance reshaping occurred without eye movements, and it occurred when the dynamics of the image statistics mimicked naturally occurring dynamics. Third, we found that exposure to "broken" temporal contiguity image statistics could weaken and even reverse the previously normal IT object selectivity at a specific position or

size (i.e., exposure could break old correct tolerance and build new "incorrect" tolerance), and that naturally occurring temporal contiguity image statistics could build new, correct position or size tolerance. Taken together with previous work, these results argue that the ventral stream uses unsupervised, natural visual experience and a common learning mechanism (a.k.a. unsupervised temporal tolerance learning, or UTL) to build and maintain its tolerant (invariant) object representation.

RESULTS

In three separate experiments (Experiments I, II, III), two unsupervised nonhuman primates (Rhesus monkeys, *Macaca mulatta*) were exposed to altered visual worlds in which we manipulated the temporal contiguity statistics of the animals' visual experience with object size (Figure 1A, *Exposure Phases*). In each experiment, we recorded multiunit activity (MUA) in an unbiased sample of recording sites in the anterior region of IT to monitor any experience-induced change (Figure 1A, *Test Phases*). Specifically, for each IT site, a preferred object (P) and a less-preferred object (N) were chosen based on testing of a set of 96 objects (Figure 1B). We then measured the baseline IT neuronal selectivity for P and N at three retinal sizes (1.5°, 4.5°, and 9°) in a *Test Phase* (~10 min) by presenting the object images in a rapid but naturally paced sequence (5 images/s) on the animals' center of gaze. For all the results below, we report selectivity values determined from these *Test Phases*, which we conducted both before and after experience manipulations. Thus, all response data shown in the results below were collected during orthogonal behavioral tasks in which object identity and size were irrelevant (Supplemental Experimental Procedures available online).

Consistent with previous reports (Kreiman et al., 2006), the initial *Test Phase* data showed that each IT site tended to maintain its preference for object P over object N at each size tested here (Figures 3 and S3 available online). That is, most IT sites showed good, baseline size tolerance. Following the logic outlined in the Introduction, the goal of Experiments I–III was to determine if consistently applied, unsupervised experience manipulations would predictably reshape that baseline size tolerance of each IT site (see Figure 1 for the basic prediction). In particular, we monitored changes in each IT site's preference for object P over N at each of the three objects sizes, and any change in that selectivity following experience that was not seen in control conditions was taken as evidence for an experience-induced reshaping of IT size tolerance.

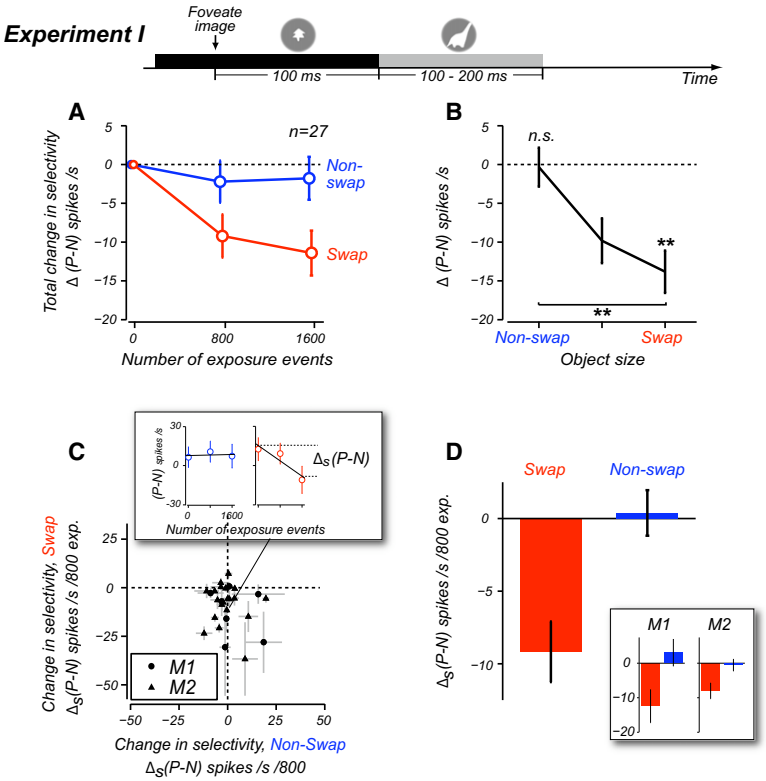
In each experiment, the key experience manipulation was deployed in one or more *Exposure Phases* that were all under precise, automated computer-display control to implement spatiotemporally reliable experience manipulations (see Experimental Procedures). Specifically, during each *Exposure Phase* the animals freely viewed a gray display monitor on which images of object P or N intermittently appeared at randomly chosen retinal positions away from the center of gaze (object size: 1.5°, 4.5°, or 9°). The animals almost always looked to foveate each object (>95% of object appearances) within ~124 ms (mean; median, 109 ms), placing the object image on the center of gaze. Following that object acquisition saccade, we reliably manipulated the visual experience of the animals over the next 200–300 ms. The details of the experience manipulation (i.e., which object sizes where shown and the timing of those object images) were different in the three experiments, but all three experiments used the same basic logic outlined in the Introduction and in Figure 1.

Experiment I: Does Unsupervised Visual Experience Reshape IT Size Tolerance?

In Experiment I, following the object acquisition saccade, we left the newly foveated object image unchanged for 100 ms, and then we changed the size of the object image (while its retinal position remained on the animal's center of gaze) for the next 100 ms (Figure 1A). We reasoned that this creates a temporal experience linkage ("exposure event") between one object image at one size and another object image at another size. Importantly, on half of the exposure events, one object was swapped out for the other object: for example, a medium-sized (4.5°) object P would become a big (9°) object N (Figure 1A, "swap exposure event"). As one key control, we also exposed the animal to more normal exposure events in which object identity did not change during the size change (Figure 1A, "non-swap exposure event"). The full exposure design for one IT site is shown in Figure 1B; the animal received 800–1600 swap exposures within the time period of 2–3 hr. Each day, we made continuous recordings from a single IT site, and we always deployed the swap exposure at a particular object size (either 1.5° or 9°, i.e., swap size) while keeping the other size as a control (i.e., non-swap size). Across different IT sites (i.e., different recording days), we strictly alternated the object size at which swap manipulation took place so that object size was counterbalanced across our recorded IT population (n = 27).

UTL theory makes the qualitative prediction that the altered experience will induce a size-specific confusion of object identity in the IT response as the ventral stream learns to associate the temporally linked images. In particular, our exposure design should cause the IT site to reduce its original selectivity for images of object P and N at the swap size (perhaps even reversing that selectivity in the limit of large amounts of experience; Figure 1C, red). UTL is not currently specific enough to make a quantitative prediction of what this altered experience should do for selectivity among the medium object size images because those images were temporally paired in two ways: with images at the swap size (altered visual experience) and with the images at the non-swap size (normal visual experience). Thus, our key experimental prediction and planned comparison is between the selectivity (P versus N) at the swap and non-swap size: we predict a selectivity decrease at the swap size that should be much larger than any selectivity change at the non-swap object size (Figure 1C, blue).

This key prediction was born out by the data: as the animals received experience in the altered visual world, IT selectivity among objects P and N began to decrease at the swap size, but not at the control size. This change in selectivity grew stronger with increasing experience over the time course of 2–3 hr (Figure 2A). To quantify the selectivity change, for each IT site, we took the difference between the selectivity (P − N, response difference in units of spikes/s, see Experimental



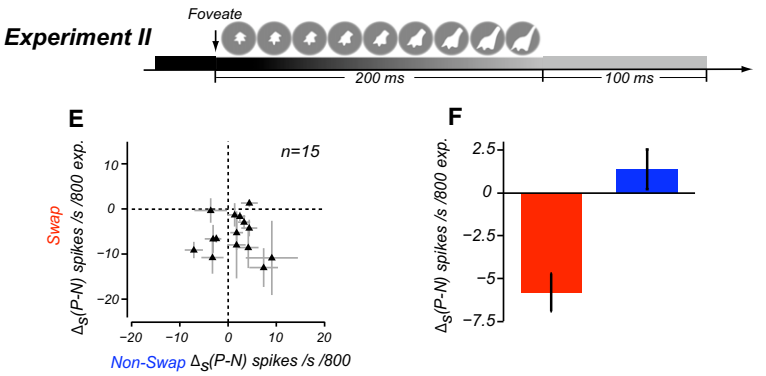


Figure 2. Experimental I and II Key Results

(A) Mean ± SEM. IT object selectivity change, Δ(P − N), from the first *Test Phase* as a function of the number of exposure events is shown. Each data point shows the average across all the sites tested for that particular amount of experience (n = 27, 800 exposure events; n = 22, 1600 exposure events).

(B) Mean ± SEM selectivity change at the swap, non-swap, and medium size (4.5°). For each IT site (n = 27), total Δ(P − N) was computed using the data from the first and last *Test Phase*, excluding any middle *Test Phase* data. Hence, not all data from (A) were included. *p < 0.05 by two-tailed t test; **p < 0.01; n.s. p > 0.05.

(C) For each IT site (n = 27), we fit a line (linear regression) to the (P − N) data as a function of the number of exposure events (insert). We used the slope of the line fit, Δs(P − N), to quantify the selectivity change. The Δs(P − N) is a measure that leverages all our data while normalizing out the variable of exposure amount [for sites with only two *Test Phases*, Δs(P − N) equals Δ(P − N)]. Δs(P − N) was normalized to show selectivity change per 800 exposure events. Error bars indicate the standard error of the procedure to compute selectivity (Supplemental Experimental Procedures). M1, monkey 1; M2, monkey 2.

(D) Mean Δs(P − N) at the swap and non-swap size (n = 27 IT sites; M1: 7, M2: 20). Error bars indicate SEM over neuronal sites.

(E) Change in selectivity, Δs(P − N), of all IT sites from Experiment II at the swap and non-swap size.

(F) Mean ± SEM Δs(P − N) at the swap and non-swap size.

Procedures) in the first (pre-exposure) and last *Test Phase*. This Δ(P − N) sought to quantify the total amount of selectivity change for each IT site induced by our experience manipulation. On average, there was a significant decrease in selectivity at the swap size (Figure 2B, p < 0.0001, two-tailed t test against 0) and no significant change at the non-swap control size (Figure 2B, p = 0.89). Incidentally, we also observed a significant decrease in selectivity at the medium size (p = 0.002). This is not surprising given that the images at the medium object size were exposed to the altered statistics half of the time when they were temporally paired with the images at the swap size. Because no prediction was made about the selectivity change at the medium size, we concentrate below on the planned comparison between the swap and non-swap size. We statistically confirmed the size specificity of the experience-induced decrease in selectivity by two different approaches: (1) a direct t test on the Δ(P − N) between the swap and non-swap size (p < 0.001, two-tailed), and (2) a significant interaction of "exposure × object size" on the raw selectivity measurements (P − N)—that is, IT selectivity was decreased by exposure only at the swap size (p = 0.0018, repeated-measures ANOVA; p = 0.006, bootstrap, see Supplemental Experimental Procedures).

To ask if the experience-induced selectivity change was specific to the manipulated objects or the features contained in those objects, we also tested each IT site's responses to a second pair of objects (P′ and N′, control objects; see Experimental Procedures). Images of these control objects at three sizes were tested together with the swap objects during all *Test Phases* (randomly interleaved), but they were not shown during the *Exposure Phase*. On average, we observed no change in IT selectivity among these unexposed control objects (Figure S4). This shows that that the experience-induced reshaping of IT size tolerance has at least some specificity for the experienced objects or the features contained in those objects.

We next set out to quantify the amount of IT size tolerance reshaping induced by the altered visual experience. Because each IT site was tested for different amounts of exposure time (due to experimental time constraints), we wanted to control for this and still leverage all the data for each site to gain maximal power. To do so, we fit linear regressions to the (P − N) selectivity of individual sites at each object size (Figure 2C, insert). The

slope of the line fit, which we will refer to as Δs(P − N), provided us with a sensitive, unbiased measure of the amount of selectivity change that normalizes the amount of exposure experience. The Δs(P − N) for the swap size and non-swap size is shown in Figures 2C and 2D, which qualitatively confirmed the result obtained in Figure 2B (using the simple measure of selectivity change), and showed a mean selectivity change of −9.2 spikes/s for every 800 swap exposure events.

Importantly, we note that this reshaping of IT tolerance was induced by unsupervised exposure to temporally linked images that did not include a saccadic eye movement to make that link (Figure 1A). We also considered the possibility that small intervening microsaccades might still have been present, but found that they cannot account for the reshaping (Figure S7). The size specificity of the selectivity change also rules out alternative explanations such as adaptation, which would not predict this specificity (because our exposure design equated the amount of exposure for both the swap and non-swap size). We also found the same amount of tolerance reshaping when the sites were grouped by the physical object size at which we deployed the swap (1.5° versus 9°, p = 0.26, t test). Thus the learning is independent of low-level factors like the total luminance of the swapped objects. In sum, we found that unsupervised, temporally linked experience with object images across object size change can reshape IT size tolerance.

Experiment II: Does Size Tolerance Learning Generalize to the "Natural" Visual World?

In the natural world, objects tend to undergo size change smoothly on our retinas as a result of object motion or viewer motion, but, in Experiment I (above), the object size changes we deployed were discontinuous: one image of an object was immediately replaced by an image of another object with no smooth transition (Figure 2, top). Therefore, although those results show that unsupervised experience with object images at different sizes linked in time could induce the predicted IT selectivity change, we wanted to know if that learning was also found during exposure to more natural (i.e., temporally smooth) image dynamics.

To answer this question, we carried out a second experiment (Experiment II) in which we deployed essentially the same manipulation as Experiment I (object identity changes during object size changes, no intervening eye movement), but with natural (i.e., smooth-varying) stimulus sequences. The dynamics in these movie stimuli were closely modeled after the kind of dynamics that our visual system encounters daily in the natural environment (Figure S2). To create smooth-varying object identity changes over object size changes, we created morph lines between pairs of objects we swapped in Experiment I (P and N). This allowed us to parametrically transform the shape of the objects (Figure 2, bottom). All other experimental procedures were identical to Experiment I except, in the *Exposure Phases*, objects underwent size change smoothly while changing identity (swap exposure) or preserving identity (non-swap exposure, Figure S2).

When we carried out this temporally smooth experience manipulation on a new population of IT sites (n = 15), we replicated the Experiment I results (Figures 2E and 2F): there was a predicted decrease in IT selectivity at the swap size and not at the non-swap control size. This size specificity of the effect was, again, confirmed statistically by (1) direct t test on the total selectivity change, Δ(P − N), between the swap and non-swap size [Δ(P − N) = −10.3 spikes/s at swap size, +2.8 at non-swap size; p < 0.0001, two-tailed t test]; and (2) a significant interaction of "exposure × object size" on the raw selectivity measurements (P − N) (p < 0.001, repeated-measures ANOVA; p = 0.001, bootstrap). This result suggests that image linking across time is sufficient to induce tolerance learning in IT and is robust to the temporal details of that image linking (at least over the ~200 ms time windows of linking used here). More importantly, Experiment II shows that unsupervised size tolerance learning occurs in a spatiotemporal image regime encountered in real-world vision.

Size Tolerance Learning: Observations and Effect Size Comparison

Despite a wide diversity in the initial tuning of the recorded IT multiunit sites, our experience manipulation induced a predictable selectivity change that was large enough to be observed in individual IT sites: 40% (17/42 sites, Experiment I and II data combined) of the individual IT sites showed a significant selectivity decrease at the swap size within a single recording session (only 7% of sites showed significant selectivity decrease at the non-swap size, which is essentially the fraction expected by chance; 3/42 sites, p < 0.05, permutation test, see Supplemental Experimental Procedures). Eight example sites are shown in Figure 3.

We found that the magnitude of size-tolerance reshaping depended on the initial selectivity at the medium object size, 4.5° (Pearson correlation, r = 0.54, p < 0.01). That is, on average, IT sites that we initially encountered with greater object selectivity at the medium size underwent greater exposure-induced selectivity change at the swap size. This correlation is not simply explained by the hypothesis that it is easier to break highly selective neurons (e.g., due to factors that might have nothing to do with neuronal learning, such as loss of isolation), because the correlation was not seen for changes in selectivity at the non-swapped size (r = −0.16, p = 0.35) and we found no average change in selectivity at the non-swapped size (Figure 2 and statistics above). Instead, this observation is consistent with the overarching hypothesis of this study: the initial image selectivity at the medium object size provides (at least part of) the driving force for selectivity learning because those images are temporally linked with the swapped images at the swap size.

The change in selectivity produced by the experience manipulation was found throughout the entire time period of the IT response, including the earliest part of that period where IT neurons are just beginning to respond above baseline (~100 ms from stimulus onset, Figure S5). This shows that the experience-induced change in IT selectivity cannot be explained by changes in long lag feedback alone (>100 ms; also see Discussion). On average, the selectivity change at the swap size resulted from both a decrease in the response to the image of the preferred object (P) and an increase in the response to the less preferred object (N). Consistent with this, we found that the

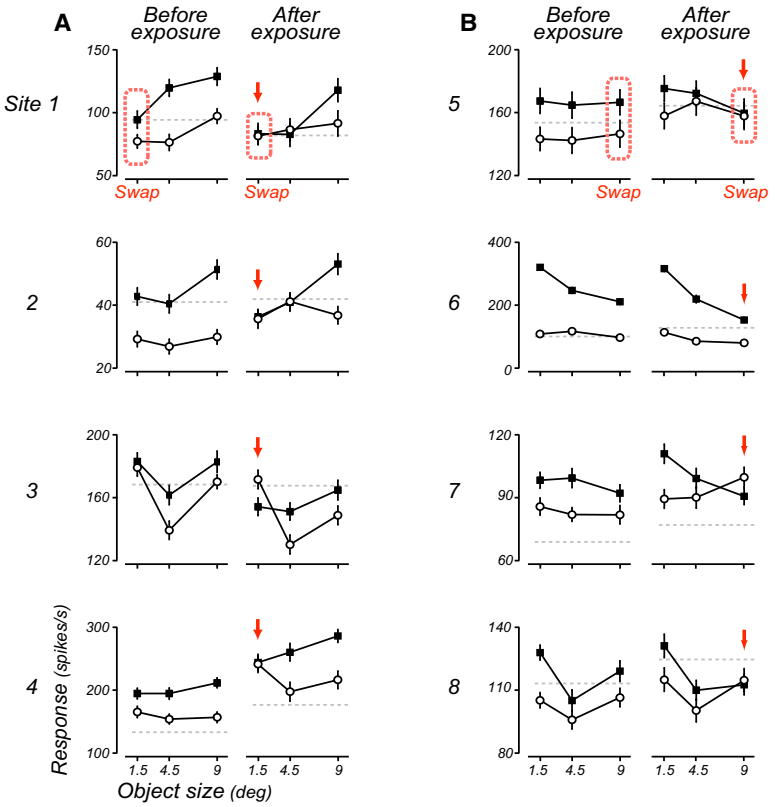


Figure 3. Example Single IT Sites

Mean ± SEM. IT response to P (solid square) and N (open circle) as a function of object size for eight example IT sites (from both Experiment I and II). The data shown are from the first ("before exposure") and last ("after exposure") *Test Phase*. (A) Swap size, 1.5°; (B) swap size, 9° (highlighted by red boxes and arrows). Gray dotted lines show the baseline response to a blank image (interleaved with the test images).

experience manipulation produced no average change in the IT sites' mean response rate (Figure S5).

In this study, we concentrated on multiunit response data because it had a clear advantage as a direct test of our hypothesis—it allowed us to longitudinally track IT selectivity during altered visual experience across the entirety of each experimental session. We also examined the underlying single-unit data and found results that were consistent with the multiunit data. Figure 4A shows an example of a rare single-unit IT neuronal recording that we were able to track across an entire recording session (~3 hr). The confidence that we were recording from the same unit comes from the consistency of the unit's waveform and its consistent pattern of response among the nonexposed control object images (Figure 4B). During this stable recording, the (P − N) selectivity at the swap size gradually decreased while the selectivity at the non-swap size remained stable, perfectly mirroring the multiunit results described above. However these ~3 hr single-unit recordings were very rare because single units have limited hold-time in the awake primate physiology preparation. Thus we took a more standard population approach to analyze the single-unit data (Baker et al., 2002; Kobatake et al., 1998; Sakai and Miyashita, 1991; Sigala et al., 2002). Specifically, we performed spike-sorting analyses to obtain clear single units from each *Test Phase* (Experimental Procedures). We considered each single unit obtained from each *Test Phase* as a sample of the IT population, taken either before or after the experience in the altered visual world. This analysis does not require that the sampled units were the same neurons. The prediction is that IT single units sampled after exposure (i.e., at the last *Test Phase* of each day) would be less size tolerant at the swap size than at the non-swap size. This prediction was clearly observed in our single-unit data (Figure 4C, after exposure, $p < 0.05$; for reference, the size tolerance before the exposure is also shown and we observed no difference between the swap and non-swap size). The result was robust to the choice of the criteria to define "single units" (Figure S6). Similarly, we found that each single-unit population sampled after successively more exposure showed a successively larger change in size tolerance (Figure 4D).

We next aimed to quantify the absolute magnitude of this size tolerance learning effect across the different experience manipulations deployed here, and to compare that magnitude with our previous results on position-tolerance learning (Li and DiCarlo, 2008). To do this, we plotted the mean selectivity change at the swap size from each experiment as a function of number of swap exposures (Figure 5). We found that Experiments I and II produced a very similar magnitude of learning: ~5 spikes/s per 400 swap exposures (also see Discussion for comparison to previous work). This effect grew larger at this approximately constant rate for as long as we could run each experiment, and the magnitude of the size tolerance learning was remarkably similar to that seen in our previous study of position tolerance (Li and DiCarlo, 2008).

Size and Position Tolerance Learning: Reversing Old IT Object Selectivity and Building New IT Object Selectivity

The results on size tolerance presented above and our previous study of position tolerance (Li and DiCarlo, 2008) both used the breaking of naturally occurring temporal contiguity experience to discover that we can break normal position tolerance and size tolerance (i.e., we can cause a decrease in adult IT object selectivity in a size- or position-specific manner). While these results are consistent with the inference that naturally occurring image statistics instruct the original building of that normal tolerance (see Introduction), we next sought to test that inference more directly. Specifically, we asked if the temporal contiguity statistics of visual experience can instruct the creation of new IT tolerance (i.e., if they can cause an increase in IT object selectivity in a size- or position-specific manner). Our experimental data offered two ways to test this idea (below), and both ways revealed that unsupervised temporal contiguity learning could

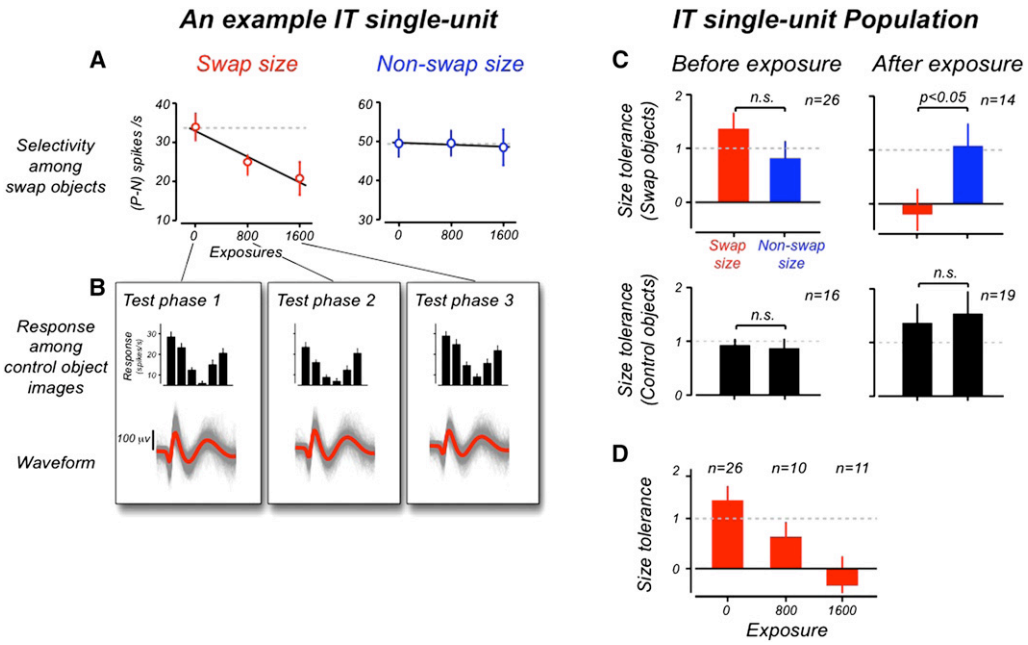


Figure 4. Single-Unit Results

(A) P versus N selectivity of a rare single-unit IT neuron that was isolated across an entire recording session (~3 hr).

(B) The example single-unit's response to the six control object images during each *Test Phase* and its waveforms (gray: all traces from a *Test Phase*; red: mean).

(C) Mean ± SEM size tolerance at the swap (red) and non-swap (blue) size for single units obtained before and after exposure. Size tolerance for the control objects is also shown at these two sizes (black). Each neuron's size tolerance was computed as $(P - N)/(P - N)_{medium}$, where $(P - N)$ is the selectivity at the tested size and $(P - N)_{medium}$ is the selectivity at the medium object size. Only units that showed selectivity at the medium size were included [$(P - N)_{medium} > 1$ spikes/s]. The top and bottom panels include neurons that had selectivity for the swap objects, the control objects, or both. Thus they show different but overlapping populations of neurons. The result is unchanged if we only examine populations for which each neuron has selectivity for both the swap and control objects (i.e., the intersections of the neuronal populations in top and bottom panels; Figure S6).

(D) Mean ± SEM size tolerance at the swap size further broken out by the amount of exposure to the altered visual statistics. To quantify the change in IT size tolerance, we performed linear regression of the size tolerance as a function of the amount of experience. Consistent with the multiunit results, we found a significant negative slope (Δ size tolerance = −0.84 per 800 exposure; p = 0.002, bootstrap; c.f. −0.42 for multiunit, Figure S6). No decrease in size tolerance was observed at the non-swap control size (Δ size tolerance = 0.30; c.f. 0.12 for multiunit).

indeed build new IT tolerance. To do these analyses, we took advantage of the fact that we found very similar effects for both size tolerance and position tolerance (Li and DiCarlo, 2008), and we maximized our power by pooling the data across this experiment (Figure 5: size experiment I, II; n = 42 MUA sites) and our previous position experiment (n = 10 MUA sites). This pooling did not qualitatively change the result—the effects shown in Figures 5 and 6 below were seen in the size tolerance data alone (Figure S9).

First, as outlined in Figure 1C, a strong form of the UTL hypothesis predicts that our experience manipulation should not only degrade existing IT selectivity for P over N at the swap size/position, but should eventually reverse that selectivity and then build new incorrect selectivity for N over P (Figure 1C; note that we refer to this as incorrect selectivity because the full IT response pattern is inappropriate for the veridical world in which objects maintain their identity across changes in position and size). Though the plasticity we discovered is remarkably strong (~5 spikes/s per hour), it did not produce a selectivity reversal for the "mean" IT site within the 2 hr recording session (Figure S5D). Instead, it only produced a ~50% decrease in selectivity for that mean site, which is entirely consistent with the fact that our mean IT site had reasonably strong initial selectivity for P over N (mean P − N = ~20 spikes/s). To look more deeply at this issue, we made use of the well-known observation that not all adult IT neurons are identical— some have a large amount of size or position tolerance, whereas others show a small amount of tolerance (DiCarlo and Maunsell, 2003; Ito et al., 1995; Logothetis and Sheinberg, 1996; Op De Beeck and Vogels, 2000). Specifically, some IT sites strongly prefer object P to N at some sizes/positions, but show only weak (P − N) selectivity at the swap sizes/positions (this neuronal response pattern is illustrated schematically at the top of Figure 6). We reasoned that examination of these sites should reveal whether our experience manipulation is capable of causing a reversal in selectivity and building of new selectivity. Thus, we used independent data to select neuronal subpopulations from our data pool with varying amounts of initial selectivity at the swap size/position (Supplemental Experimental Procedures). Note that all of these neuronal sites had robust selectivity for P over N at the medium sizes/positions (as schematically illustrated in Figure 6A). This analysis revealed that our manipulation caused neuronal sites with weak initial selectivity at the swap size/position to reverse their selectivity, and to build new

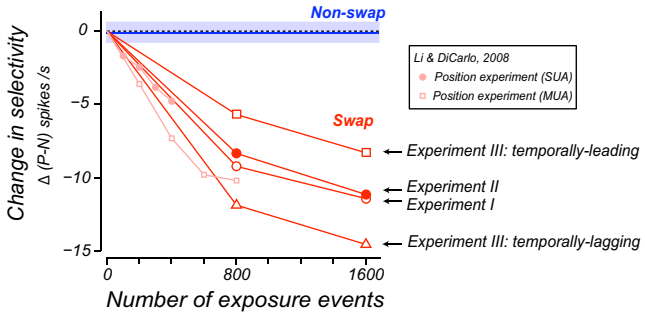


Figure 5. Effect Size Comparisons across Different Experience Manipulations

Mean object selectivity change as a function of the number of swap exposure events for different experiments. For comparison, the data from a position tolerance learning experiment (Li and DiCarlo, 2008) are also shown. Plot format is the same as Figure 2A without the error bars. Mean ± SEM Δ (P − N) at the non-swap size/position is shown in blue (all experiments pooled). SUA, single-unit activity; MUA, multiunit activity.

selectivity (building incorrect selectivity for N over P), exactly as predicted by the UTL hypothesis (Figure 6).

A second way in which our data might reveal whether UTL can build tolerance is to carefully look for any changes in selectivity at the non-swap (control) size/position. Our experiment was designed to present a large number of normal temporal contiguity exposures at that control size/position so that we would perfectly equate its amount of retinal exposure with that provided at the swap size/position. Although some forms of unsupervised temporal contiguity theory might predict that these normal temporal contiguity exposures should increase the (P − N) selectivity at the control size/position, we did not initially make that prediction (Figure 1C, blue) because we reasoned that most IT sites would already have strong, adult-like selectivity for object P versus N at that size/position, such that further supporting statistics would have little to teach those IT sites (Figure 7A, top right). Consistent with this, we found little mean change in (P − N) selectivity for the control condition in either our position tolerance experiment (Li and DiCarlo, 2008) or our size tolerance experiment (Figure 2, blue). However, examination of all of our IT sites revealed that some sites happened to have initially weak (P − N) selectivity at the control size/position while still having strong selectivity at the medium size/position (Figure 7A, top left). This suggested that these sites might be in a more naive state with respect to the particular objects being tested such that our temporal contiguity statistics might expand their tolerance for these objects (i.e., increase their P − N selectivity at the control size/position). Indeed, examination of these sites reveals that our exposure experiment caused a clear, significant building of new, correct selectivity among these sites (Figure 7B), again directly demonstrating that unsupervised temporal contiguity experience can build IT tolerance.

Experiment III: Does the Learning Depend on the Temporal Direction of the Experience?

Our results show that targeted alteration of unsupervised natural visual experience rapidly reshapes IT size tolerance—as predicted by the hypothesis that the ventral stream uses a temporal contiguity learning strategy to build that tolerance in the first place. Several instantiated computational models show how this conceptual strategy can build tolerance (Foldiak, 1991; Masquelier et al., 2007; Masquelier and Thorpe, 2007; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002; Wyss et al., 2006), and such models can be implemented using variants of Hebbian-like learning rules that are dependent on the timing of spikes (Gerstner et al., 1996; Sprekeler et al., 2007; Wallis and Rolls, 1997; Morrison et al., 2008; Sprekeler and Gerstner, 2009). The time course and task independence of the observed learning are consistent with synaptic plasticity (Markram et al., 1997; Meliza and Dan, 2006), but our data do not constrain the underlying mechanism. One can imagine ventral stream neurons using almost temporally coincident activity to learn which sets of its afferents correspond to features of the same object across size changes. If tolerance learning is spike timing dependent, any experience-induced change in IT selectivity might reflect any temporal asymmetries at the level of the underlying synaptic learning mechanism. For example, one hypothesis is that lingering postsynaptic activity caused by temporally leading images drives synaptic plasticity in afferents activated by temporally lagging images. Alternatively, afferents activated by temporally leading images might be modified by the later arrival of postsynaptic activity caused by temporally lagging images. Or a combination of both hypotheses might be the case. To look for reflections of any such underlying temporal asymmetry, we carried out a third experiment (Experiment III) centered on the question, "Do temporally leading images teach temporally lagging ones, or vice-versa?"

We deployed the same experience manipulation as before (linking of different object images across size changes, the same as Experiment I), but this time only in one direction (compare single-headed arrows in Figure 8A with double-headed arrows in Figure 1B). For example, during the recording of a particular IT site, the animal only received experience seeing objects temporally transition from a small size (arrow "tail" in Figure 8A) to a large size (arrow "head" in Figure 8A) while swapping identity. We strictly alternated the temporal direction of the experience across different IT sites. That is, for the next IT site we recorded, the animal experienced objects transitioning from a large size to a small size while swapping identity. Thus, object size was counterbalanced across our recorded population, so that we could isolate changes in selectivity among the temporally leading stimuli (i.e., arrow tail stimuli) from changes in selectivity among the temporally lagging stimuli (i.e., arrow head stimuli). As in Experiments I and II, we measured the expression of any experience-induced learning by looking for any change in (P − N) selectivity at each object size measured in a neutral task with all images randomly interleaved (*Test Phase*). We replicated the results in Experiments I and II in that a decrease in (P − N) selectivity was found following swapped experience (red bars are negative in Figure 8B). When we sorted our data based on the temporal direction of the animals' experience, we found greater selectivity change (i.e., learning) for the temporally lagging images (Figure 8B). This difference was statistically significant ($p = 0.038$, $n = 31$, two-tailed t test) and cannot be explained by any differences in the IT sites' initial selectivity (Figure S4C; also see Figure S4B for results with all sites

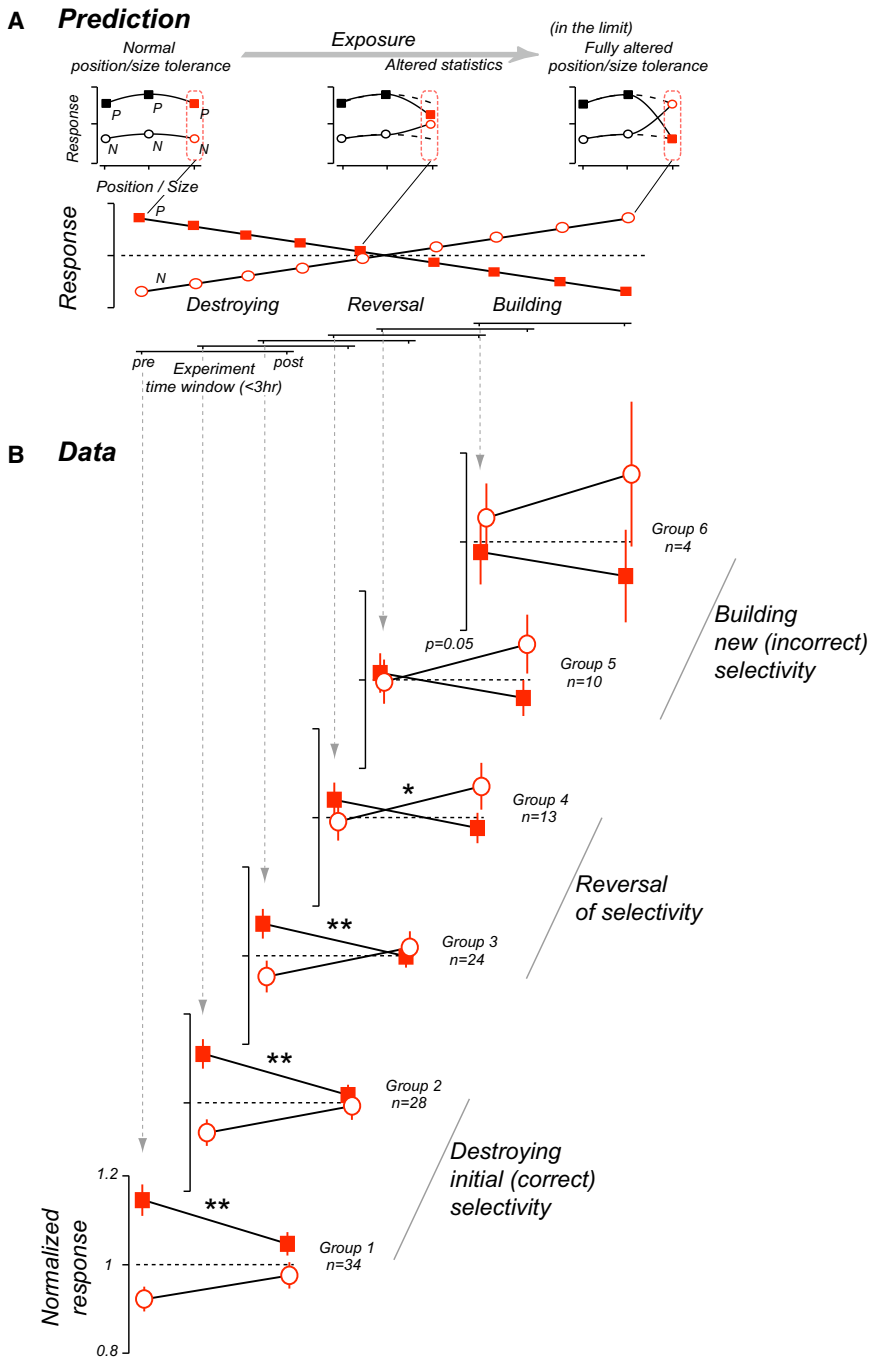


Figure 6. Altered Statistics in Visual Experience Builds Incorrect Selectivity

(A) Prediction: *top*, most adult IT neurons start with fully position/size tolerant selectivity (left). In the limit of a large amount of altered visual experience, temporal contiguity learning predicts that each neuron will acquire fully altered tolerance (right). *Bottom*, at the swap position/size (red), the selectivity for P over N is predicted to reverse in the limit (prefer N over P). Because we could only record longitudinally from a multiunit site for less than 3 hr, we do not expect our experience manipulation within a session to produce the full selectivity reversal (pre versus post) among neuronal sites with strong initial selectivity. However, because different IT sites differ in their degrees of initial selectivity, they start at different distances from selectivity reversal. Thus, our manipulation should produce selectivity reversal among the initially weakly selective sites and build new ("incorrect") selectivity.

(B) Mean ± SEM normalized response to object P and N at the swap position/size among subpopulations of IT multiunit sites. Sites are grouped by their initial selectivity at the swap position/size using independent data. Data from the size and position tolerance experiments (Li and DiCarlo, 2008) were combined to gain maximal power (size experiment I, II; position experiment, see Supplemental Experimental Procedures). These sites show strong selectivity at the non-swap (control) position/size, and no negative change in that selectivity was observed (not shown). $^{**}p < 0.01$; $^{*}p < 0.05$, one-tailed t test against no change. (Size experiment data only, group 1–6: $p < 0.01$; $p < 0.01$; $p < 0.01$; $p = 0.02$; $p = 0.07$; n.s.).

included). This result is consistent with an underlying learning mechanism that favors experience-induced plasticity of the afferents corresponding to temporally lagging images.

To test if the tolerance learning spread beyond the specifically experienced images, here, we also tested object images at an intermediate size (3°) between the two exposed sizes (Figure 8). Unlike as in Experiments I and II, this medium size was not exposed to the animals during the *Exposure Phase* (it was also at a different physical size from the medium size in Experiments I and II). We observed significant selectivity change for the medium size image pairs (Figure 8B, middle bar; $p = 0.01$, two-tailed t test against zero), which suggests that the tolerance learning has some degree of spread (but not to very different objects; Figure S4). Finally, the effect size observed in Experiment III was consistent with, and can explain the effect sizes observed in Experiments I and II. That is, based on the Experiment III effect sizes for the temporally lagging and leading images, a first-order prediction of the net effect in Experiments I and II is the average of these two effects (because Experiments I and II employed a 50-50 mix of the experience manipulations considered separately in Experiment III). That prediction is very close to what we found (Figure 5).

DISCUSSION

The overarching goal of this work is to ask whether the primate ventral visual stream uses a general, temporal contiguity driven learning mechanism to construct its tolerance to

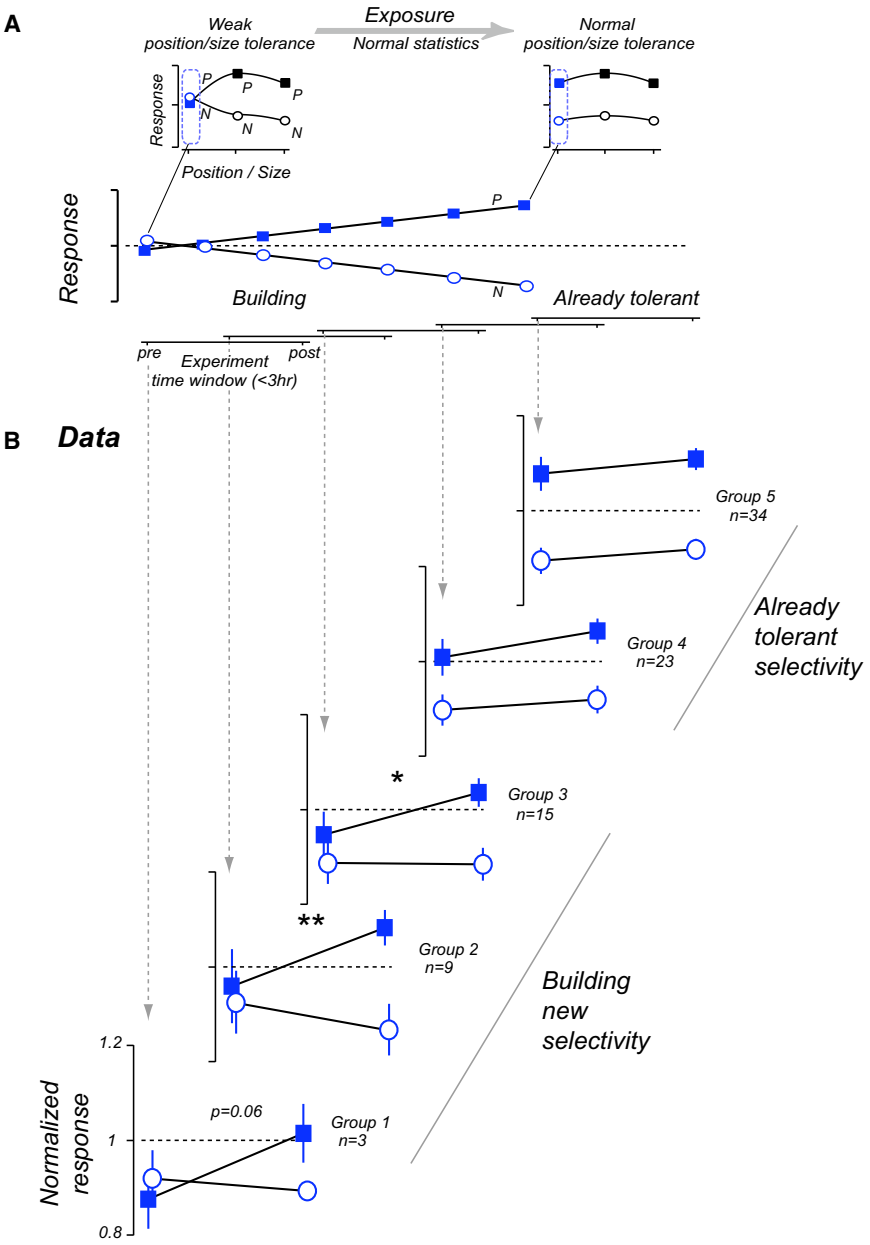


Figure 7. Normal ("Correct") Statistics in Visual Experience Builds Tolerant Selectivity

(A) Prediction follows the same logic as in Figure 6A, but here for the control conditions in which normal temporal contiguity statistics were provided (Figure 1). *Top*, temporal contiguity learning predicts that neurons will be taught to build new "correct" selectivity (i.e., normal tolerance), and neurons starting with initially weak position/size tolerant selectivity (left) have the highest potential to reveal that effect. *Bottom*, at the non-swap position/size (blue), our manipulation should build new correct selectivity for P over N among IT sites with weak initial selectivity. (B) Mean ± SEM normalized response to object P and N at the non-swap position/size among subpopulations of IT multiunit sites. Sites are grouped by their initial selectivity at the non-swap position/size using independent data. Other details are the same as those in Figure 6B. (Size experiment data only, group 1–5: $p = 0.06$; $p < 0.01$; $p = 0.05$; n.s.; n.s.).

object-identity-preserving image transformations. Our strategy was to use experience manipulations of temporally contiguous image statistics to look for changes in IT neuronal tolerance that are predicted by this hypothetical learning mechanism. Here we tested three key predictions that were not answered by previous work (Li and DiCarlo, 2008). First, we asked if these experience manipulations predictably reshaped the size tolerance of IT neurons. Our results strongly confirmed this prediction: we found that the change in size tolerance was large (~5 spikes/s, ~25% IT selectivity change per hour of exposure) and grew gradually stronger with increasing visual experience. Second, we asked if this tolerance reshaping was induced under visual experience that mimics the common size-tolerance-building statistics in the natural world: temporally contiguous image changes without intervening eye movements, and temporally smooth dynamics. Our results confirmed this prediction: we found that size tolerance was robustly reshaped in both of these conditions (Figure 2), and the magnitude of reshaping was similar to that seen with eye-movement-contingent reshaping of IT position tolerance (Li and DiCarlo, 2008, Figure 5). Third, we asked if experience with temporal contiguous image statistics could not only break existing IT tolerance, but could also build new tolerance. Again, our results confirmed this prediction: we found that experience with incorrect statistics can build incorrect tolerance (Figure 6) and that experience with correct statistics can build correct tolerance (Figure 7). Finally, we found that this tolerance learning is temporally asymmetric and spreads beyond the specifically experienced images (Figure 8, medium size), results that have implications for underlying mechanisms (see below).

Given these results, it is now highly likely that our previously reported results on eye-movement-contingent tolerance learning (Li and DiCarlo, 2008) were only one instance of a general tolerance learning mechanism. Taken together, our two studies show that unsupervised, temporally contiguous experience can reshape and build at least two types of IT tolerance, and that they can do so under a wide range of spatiotemporal regimes encountered during natural visual exploration. In sum, we speculate that these studies are both pointing to the same general learning mechanism that builds adult IT tolerance,

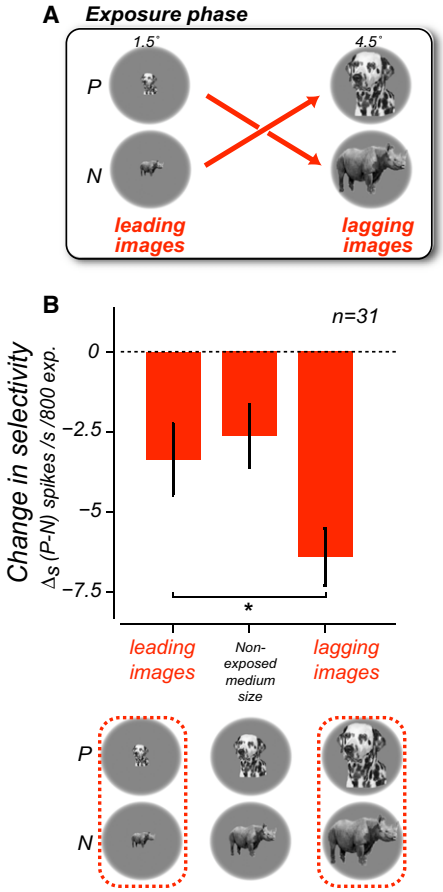


Figure 8. Experiment III Exposure Design and Key Results

(A) *Exposure Phase* design (*top*, same format as in Figure 1B) and example object images used (*bottom*).

(B) Mean ± SEM selectivity change, Δs(P − N), among the temporally leading images, the nonexposed images at the medium object size (3°), and the temporally lagging images. Δs(P − N) was normalized to show selectivity change per 800 exposure events. *p = 0.038, two-tailed t test.

and we have previously termed this mechanism "unsupervised temporal slowness learning" (Li and DiCarlo, 2008).

Our suggestion that UTL is a general tolerance learning mechanism is supported by a number of empirical commonalities between the size tolerance learning here and our previously reported position tolerance learning (Li and DiCarlo, 2008). (1) Object specificity: the experience-induced changes in IT size tolerance and position tolerance have at least some specificity for the exposed object. (2) Learning induction (driving force): in both studies, the magnitude of learning depended on the initial selectivity of the temporally adjacent images (medium object size here, foveal position in the position tolerance study), which is consistent with the idea that the initial selectivity may provide at least part of the driving force for the learning. (3) Time course of learning expression: learning increased with increasing amount of experience and changed the initial part of IT response (100 ms after stimulus onset). (4) Response change of learning expression: in both studies, the IT selectivity change arose from a response decrease to the preferred object (P) and a response increase to the less preferred object (N). (5) Effect size: our different experience manipulations here as well as our previous position manipulation revealed a similar effect magnitude (~5 spikes/s per 400 swap exposures). More specifically, when measured as learning magnitude per exposure event, size tolerance learning was slightly smaller than that found for position tolerance learning (Figure 5), and when considered as learning magnitude per unit time, the results of all three experiments were nearly identical (Figure S8). However, we note that our data cannot cleanly deconfound exposure amount from exposure time.

Relation to Previous Literature

Previous psychophysical studies have shown that human object perception depends on the statistics of visual experience (e.g., Brady and Oliva, 2008; Fiser and Aslin, 2001; Turk-Browne et al., 2005). Several studies have also shown that manipulating the spatiotemporal contiguity statistics of visual experience can alter the tolerance of human object perception (Cox et al., 2005; Wallis et al., 2009; Wallis and Bülthoff, 2001). In particular, an earlier study (Cox et al., 2005) showed that the same type of experience manipulation deployed here (experience of different object images across position change) produces increased confusion of object identities across position—a result that qualitatively mirrors the neuronal results reported here and in our previous neuronal study (Li and DiCarlo, 2008). Thus, the available psychophysical data suggest that UTL has perceptual consequences. However, this remains an open empirical question (see "Limitations and Future Direction" subsection).

Previous neurophysiological investigations in the monkey ventral visual stream showed that IT and perirhinal neurons could learn to give similar responses to temporally nearby stimuli when instructed by reward (i.e., so-called "paired associate" learning; Messinger et al., 2001; Miyashita, 1988; Sakai and Miyashita, 1991), or sometimes, even in the absence of reward (Erickson and Desimone, 1999). Though these studies were motivated in the context of visual memory (Miyashita, 1993) and used visual presentation rates of seconds or more, it was recognized that the same associational learning across time might also be used to learn invariant visual features for object recognition (e.g., Foldiak, 1991; Stryker, 1991; Wallis, 1998; Wiskott and Sejnowski, 2002). Our studies provide a direct test of these ideas by showing that temporally contiguous experience with object images can specifically reshape the size and position tolerance of IT neurons' selectivity among visual objects. This is consistent with the hypothesis that the ventral visual stream relies on a temporal contiguity strategy to learn its tolerant object representations in the first place. Our results also demonstrate that UTL is somewhat specific to the experienced objects' images (i.e., object, size, position specificity) and operates over natural, very fast time scales (hundreds of ms, faster than those previously reported) in a largely unsupervised manner. This suggests that, during natural visual exploration, the visual system can leverage an enormous amount of visual experience to construct its object invariance.

Computational models of the ventral visual stream have put forms of the temporal contiguity hypothesis to test, and have shown that learning to extract slowly varying features across

time can produce tolerant feature representations with units that mimic the basic response properties of ventral stream neurons (Masquelier et al., 2007; Masquelier and Thorpe, 2007; Sprekeler et al., 2007; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002; Wyss et al., 2006). These models can be implemented using variants of Hebbian-like learning rules (Masquelier and Thorpe, 2007; Sprekeler and Gerstner, 2009; Sprekeler et al., 2007; Wallis and Rolls, 1997). The time course and task independence of UTL reported here is consistent with synaptic plasticity (Markram et al., 1997; Rolls et al., 1989), and the temporal asymmetry in learning magnitude (Figure 8) constrains the possible underlying mechanisms. While the experimental approach used here may seem to imply that experience with all possible images of each object is necessary for UTL to build an invariant IT object representation, this is not believed to be true in a full computational model of the ventral stream. For example, V1 complex cells that encode edges may learn position tolerance that ultimately supports the invariant encoding of many objects. Our observation of partial spread of tolerance learning to nonexperienced images (Figure 8) is consistent with this idea. In particular, at each level of the ventral stream, afferent input likely reflects tolerance already constructed for simpler features at the previous level (e.g., in the context of this study, some IT afferents may respond to an object's image at both the medium size and the swap size). Thus any modification of the swap-size-image-afferents would result in a partial generalization of the learning beyond the specifically experienced images.

Limitations and Future Direction

Because the change in object selectivity was *expressed* in the earliest part of the IT response after learning (Figure S5A), even while the animal was performing tasks unrelated to the object identity, this rules out any simple attentional account of the effect. However, our data do not rule out the possibility that attention or other top down signals may be required to mediate the learning during the *Exposure Phase*. These potential top-down signals could include nonspecific reward, attentional, and arousal signals. Indeed, psychophysical evidence (Seitz et al., 2009; Shibata et al., 2009) and physiological evidence (Baker et al., 2002; Freedman and Assad, 2006; Froemke et al., 2007; Goard and Dan, 2009; Law and Gold, 2008) both suggest that reward is an important factor that can modulate or gate learning. We also cannot rule out the possibility that the attentional or the arousal system may be required for the learning to occur. In our work, we sought to engage the subjects in natural exploration during the *Exposure Phases* under the assumption that visual arousal may be important for ongoing learning, even though we deployed the manipulation during the brief periods of fixation during that exploration. Future experiments in which we systematically control these variables will shed light on these questions, and will help expose the circuits that underlie UTL.

Although the UTL phenomenology induced by our experiments was a very specific change in IT neuronal selectivity, the magnitude of this learning effect was quite large when expressed in units of spikes per second (Figure 5: ~5 spikes/s, ~25% change in IT selectivity per hour of exposure). This is comparable to or larger than other important neuronal phenomenology (e.g., attention, Maunsell and Cook, 2002). However, because this effect size was evaluated from the multiunit signal, without knowledge of how many neurons we are recording from, this effect size should be interpreted with caution. Furthermore, connecting this neuronal phenomenology (i.e., change in IT image selectivity) to the larger problem of size or position tolerance at the level of the IT population or the animal's behavior is not straightforward. Quantitatively linking a neuronal effect size to behavioral effect size requires a more complete understanding of how that neuronal representation is read out to support behavior, and large effects in confusion of object identities in individual IT neurons may or may not correspond to large confusions of object identities in perception. Such questions are the target of our ongoing and future monkey studies in which one has simultaneous measures of the neuronal learning and the animal's behaviors (modeled after those such as Britten et al., 1992; Cook and Maunsell, 2002).

The rapid and unsupervised nature of UTL gives us new experimental access to understand how cortical object representations are actively maintained by the sensory environment. However, it also calls for further characterization of the time course of this learning to inform our understanding of the stability of ventral stream object representations in the face of constantly available, natural visual experience. This sets the stage for future studies on how the ventral visual stream assembles its neuronal representations at multiple cortical processing levels, particularly during early postnatal visual development, so as to achieve remarkably powerful adult object representation.

EXPERIMENTAL PROCEDURES

Animals and Surgery

Aseptic surgery was performed on two male Rhesus monkeys (8 and 6 kg) to implant a head post and a scleral search coil. After brief behavioral training (1–3 months), a second surgery was performed to place a recording chamber to reach the anterior half of the temporal lobe. All animal procedures were performed in accordance with National Institute of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

General Design

On each experimental day, we recorded from a single IT multiunit site for 2–3 hr. During that time, the animal was provided with altered visual experience in *Exposure Phases* and we made repeated measurements of the IT site's selectivity during *Test Phases* (Figure 1). The study consisted of three separate experiments (Experiments I, II, and III), which differed from each other only in the *Exposure Phase* design (described below). We focused on one pair of objects (swap objects) that the IT site was selective for (preferred object P, and nonpreferred object N, chosen using a prescreening procedure; see Supplemental Experimental Procedures).

Experiment I

Objects (P and N at 1.5°, 4.5°, or 9°) appeared at random positions on a gray computer screen and animals naturally looked to the objects. The image of the just-foveated object was replaced by an image of the other object at a different size (swap exposure event, Figure 1A) or an image of the same object at a different size (non-swap exposure event, Figure 1A). The image change was initiated 100 ms after foveation and was instantaneous (Figure 2, top). We used a fully symmetric design illustrated graphically in Figure 1B. This experience manipulation temporally linked pairs of object images (Figure 1A shows one such link) and each link could go in both directions (Figure 1B shows full design example). For each IT site, we always deployed the swap manipulation at one particular size (referred to as the swap size: 1.5° or 9°, prechosen, strictly alternated between sites), keeping the other size as the exposure-equalized control (referred to as the non-swap size).

Experiment II

All design parameters were identical to *Experiment I* except that the image changes were smooth across time (Figure 2, bottom). The image change sequence started immediately after the animal had foveated the image and the entire sequence lasted for 200 ms (Figure S2). Identity-changing morph lines were only achievable on the silhouette shapes. Only Monkey 2 was tested in Experiment II (given the stimulus class assignment).

Experiment III

We used an asymmetric design that is illustrated graphically in Figure 8A: for each IT site, we only gave the animals experience of image changes in one direction (1.5°→4.5° or vice versa, prechosen, strictly alternated between sites). The timing of the image change was identical to that in Experiment I.

Another pair of control objects (P′ and N′, not shown in the *Exposure Phase*) was also used to probe the IT site's responses in the *Test Phase*. The selectivity among the control objects served as a measure of recording stability (below). In each *Test Phase,* the swap and control objects were tested at three sizes (Experiments I and II: 1.5°, 4.5°, 9°; Experiment III: 1.5°, 3°, 4.5°) by presenting them briefly (100 ms) on the animals' center of gaze (50–60 repetitions, randomized) during orthogonal behavioral tasks in which object identity and size were irrelevant. See Supplemental Experimental Procedures for details of the task design and behavioral monitoring.

Neuronal Assays

We recorded MUA from the anterior region of IT using standard single microelectrode methods. Our previous study on IT position tolerance learning showed that we could uncover the same learning in both single-unit activity and MUA with comparable effect size (Li and DiCarlo, 2008), so here, we only recorded MUA to maximize recording time. Over a series of recording days, we sampled across IT and sites selected for all our primary analyses were required to be selective among object P and N (ANOVA, object × sizes, $p < 0.05$ for "object" main effect or interaction) and pass a stability criterion ($n = 27$ for Experiment I; 15 for Experiment II; 31 for Experiment III). We verified that the key result is robust to the choice of the stability criteria (Figure S4). See Supplemental Experimental Procedures for details of the recording procedures and site selections.

Data Analyses

All the analyses and statistical tests were done in MATLAB (Mathworks, Natick, MA) with either custom-written scripts or standard statistical packages. The IT response to each image was computed from the spike count in a 150 ms time window (100–250 ms poststimulus onset, data from *Test Phases* only). Neuronal selectivity was computed as the response difference in units of spikes/s between images of object P and N at different object sizes. To avoid any bias in this estimate of selectivity, for each IT site we define the labels P (preferred) and N by using a portion of the pre-exposure data to determine these labels, and the remaining data to compute the selectivity values reported in the text (Supplemental Experimental Procedures). In cases where neuronal response data was normalized and combined (Figures 6 and 7), each site's response from each *Test Phase* was normalized to its mean response to all object images in that *Test Phase*. The key results were evaluated statistically using a combination of t tests and interaction tests (Supplemental Experimental Procedures). For analyses presented in Figure 4, we extracted clear single units from the waveform data of each *Test Phase* using a PCA-based spike sorting algorithm (Supplemental Experimental Procedures).

SUPPLEMENTAL INFORMATION

Supplemental Information for this article includes nine figures and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.neuron.2010.08.029.

ACKNOWLEDGMENTS

We thank Professors T. Poggio, N. Kanwisher, and E. Miller and the members of our laboratory for valuable discussion and comment on this work. We also thank J. Deutsch, B. Andken, and Dr. R. Marini for technical support. This work was supported by the NIH (grant R01-EY014970 and its ARRA supplement to J.J.D., NRSA 1F31EY020057 to N.L.) and The McKnight Endowment Fund for Neuroscience.

Accepted: August 5, 2010
Published: September 22, 2010

REFERENCES

Afraz, S.R., Kiani, R., and Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. Nature *442*, 692–695.

Baker, C.I., Behrmann, M., and Olson, C.R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. Nat. Neurosci. *5*, 1210–1216.

Brady, T.F., and Oliva, A. (2008). Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. Psychol. Sci. *19*, 678–685.

Brincat, S.L., and Connor, C.E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat. Neurosci. *7*, 880–886.

Britten, K.H., Shadlen, M.N., Newsome, W.T., and Movshon, J.A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. J. Neurosci. *12*, 4745–4765.

Cook, E.P., and Maunsell, J.H.R. (2002). Attentional modulation of behavioral performance and neuronal responses in middle temporal and ventral intraparietal areas of macaque monkey. J. Neurosci. *22*, 1994–2004.

Cox, D.D., Meier, P., Oertelt, N., and DiCarlo, J.J. (2005). 'Breaking' position-invariant object recognition. Nat. Neurosci. *8*, 1145–1147.

DiCarlo, J.J., and Maunsell, J.H.R. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. J. Neurophysiol. *89*, 3264–3278.

Erickson, C.A., and Desimone, R. (1999). Responses of macaque perirhinal neurons during and after visual stimulus association learning. J. Neurosci. *19*, 10404–10416.

Fiser, J., and Aslin, R.N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychol. Sci. *12*, 499–504.

Foldiak, P. (1991). Learning invariance from transformation sequences. Neural Comput. *3*, 194–200.

Freedman, D.J., and Assad, J.A. (2006). Experience-dependent representation of visual categories in parietal cortex. Nature *443*, 85–88.

Froemke, R.C., Merzenich, M.M., and Schreiner, C.E. (2007). A synaptic memory trace for cortical receptive field plasticity. Nature *450*, 425–429.

Gerstner, W., Kempter, R., van Hemmen, J.L., and Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. Nature *383*, 76–81.

Goard, M., and Dan, Y. (2009). Basal forebrain activation enhances cortical coding of natural scenes. Nat. Neurosci. *12*, 1444–1449.

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863–866.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. J. Neurophysiol. *73*, 218–226.

Kobatake, E., Wang, G., and Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. J. Neurophysiol. *80*, 324–330.

Kreiman, G., Hung, C.P., Kraskov, A., Quiroga, R.Q., Poggio, T., and DiCarlo, J.J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron *49*, 433–445.

Law, C.T., and Gold, J.I. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. Nat. Neurosci. *11*, 505–513.

Li, N., and DiCarlo, J.J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science *321*, 1502–1507.

Li, N., Cox, D.D., Zoccolan, D., and DiCarlo, J.J. (2009). What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? J. Neurophysiol. *102*, 360–376.

Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. Annu. Rev. Neurosci. *19*, 577–621.

Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science *275*, 213–215.

Masquelier, T., and Thorpe, S.J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. PLoS Comput. Biol. *3*, e31.

Masquelier, T., Serre, T., Thorpe, S.J., and Poggio, T. (2007). Learning complex cell invariance from natural video: a plausibility proof. CBCL Paper (Cambridge, MA: Massachusetts Institute of Technology).

Maunsell, J.H.R., and Cook, E.P. (2002). The role of attention in visual processing. Philos. Trans. R. Soc. Lond. B Biol. Sci. *357*, 1063–1072.

Meliza, C.D., and Dan, Y. (2006). Receptive-field modification in rat visual cortex induced by paired visual stimulation and single-cell spiking. Neuron *49*, 183–189.

Messinger, A., Squire, L.R., Zola, S.M., and Albright, T.D. (2001). Neuronal representations of stimulus associations develop in the temporal lobe during learning. Proc. Natl. Acad. Sci. USA *98*, 12239–12244.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature *335*, 817–820.

Miyashita, Y. (1993). Inferior temporal cortex: where visual perception meets memory. Annu. Rev. Neurosci. *16*, 245–263.

Morrison, A., Diesmann, M., and Gerstner, W. (2008). Phenomenological models of synaptic plasticity based on spike timing. Biol. Cybern. *98*, 459–478.

Op De Beeck, H., and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. J. Comp. Neurol. *426*, 505–518.

Rolls, E.T., Baylis, G.C., Hasselmo, M.E., and Nalwa, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. Exp. Brain Res. *76*, 153–164.

Sakai, K., and Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. Nature *354*, 152–155.

Seitz, A.R., Kim, D., and Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. Neuron *61*, 700–707.

Shibata, K., Yamagishi, N., Ishii, S., and Kawato, M. (2009). Boosting perceptual learning by fake feedback. Vision Res. *49*, 2574–2585.

Sigala, N., Gabbiani, F., and Logothetis, N.K. (2002). Visual categorization and object representation in monkeys and humans. J. Cogn. Neurosci. *14*, 187–198.

Sprekeler, H., and Gerstner, W. (2009). Robust learning of position invariant visual representations with OFF responses (Salt Lake City: In COSYNE).

Sprekeler, H., Michaelis, C., and Wiskott, L. (2007). Slowness: an objective for spike-timing-dependent plasticity? PLoS Comput. Biol. *3*, e112.

Stryker, M.P. (1991). Neurobiology. Temporal associations. Nature *354*, 108–109.

Tanaka, K. (1996). Inferotemporal cortex and object vision. Annu. Rev. Neurosci. *19*, 109–139.

Turk-Browne, N.B., Jungé, J., and Scholl, B.J. (2005). The automaticity of visual statistical learning. J. Exp. Psychol. Gen. *134*, 552–564.

Vogels, R., and Orban, G.A. (1996). Coding of stimulus invariances by inferior temporal neurons. Prog. Brain Res. *112*, 195–211.

Wallis, G. (1998). Spatio-temporal influences at the neural level of object recognition. Network *9*, 265–278.

Wallis, G., and Bülthoff, H.H. (2001). Effects of temporal association on recognition memory. Proc. Natl. Acad. Sci. USA *98*, 4800–4804.

Wallis, G., and Rolls, E.T. (1997). Invariant face and object recognition in the visual system. Prog. Neurobiol. *51*, 167–194.

Wallis, G., Backus, B.T., Langer, M., Huebner, G., and Bülthoff, H. (2009). Learning illumination- and orientation-invariant representations of objects through temporal association. J. Vis. *9*, 6.

Wiskott, L., and Sejnowski, T.J. (2002). Slow feature analysis: unsupervised learning of invariances. Neural Comput. *14*, 715–770.

Wyss, R., König, P., and Verschure, P.F. (2006). A model of the ventral visual system based on temporal stability and local memory. PLoS Biol. *4*, e120.

Neuron, volume *67*

Supplemental Information

Unsupervised Natural Visual Experience Rapidly Reshapes Size Invariant Object Representation in Inferior Temporal Cortex

Nuo Li and James J. DiCarlo

Supplemental Figure S1

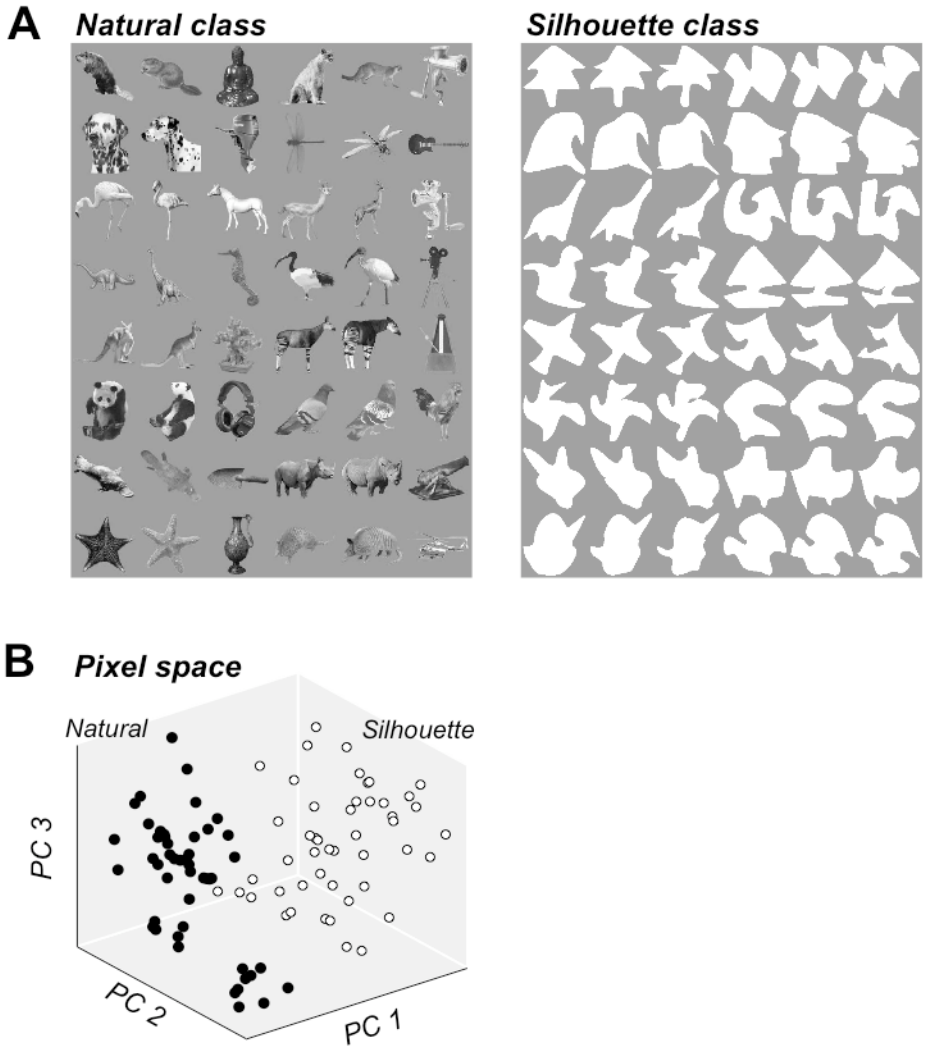


Figure S1. Stimuli and Image Analyses

(A) We selected object pairs from two different stimulus classes (48 cutout natural shapes; 48 silhouette shapes). The swap object pair (P and N used for the key experience manipulation) was always picked from one class for each animal (Monkey 1: natural; Monkey 2: silhouette). The control object pair (P' and N') was always picked from the other stimulus class.

(B) Stimuli from the two classes are quite different from each other in their pixel-wise similarity. This is illustrated when the stimulus images are plotted by scores of the first three principle components (PC) in the pixel space. Principle components were computed from all 96 images. Images were pre-processed to have equal mean and unit variance before image analyses. Solid symbols: natural; open symbols: silhouette.

Supplemental Figure S2

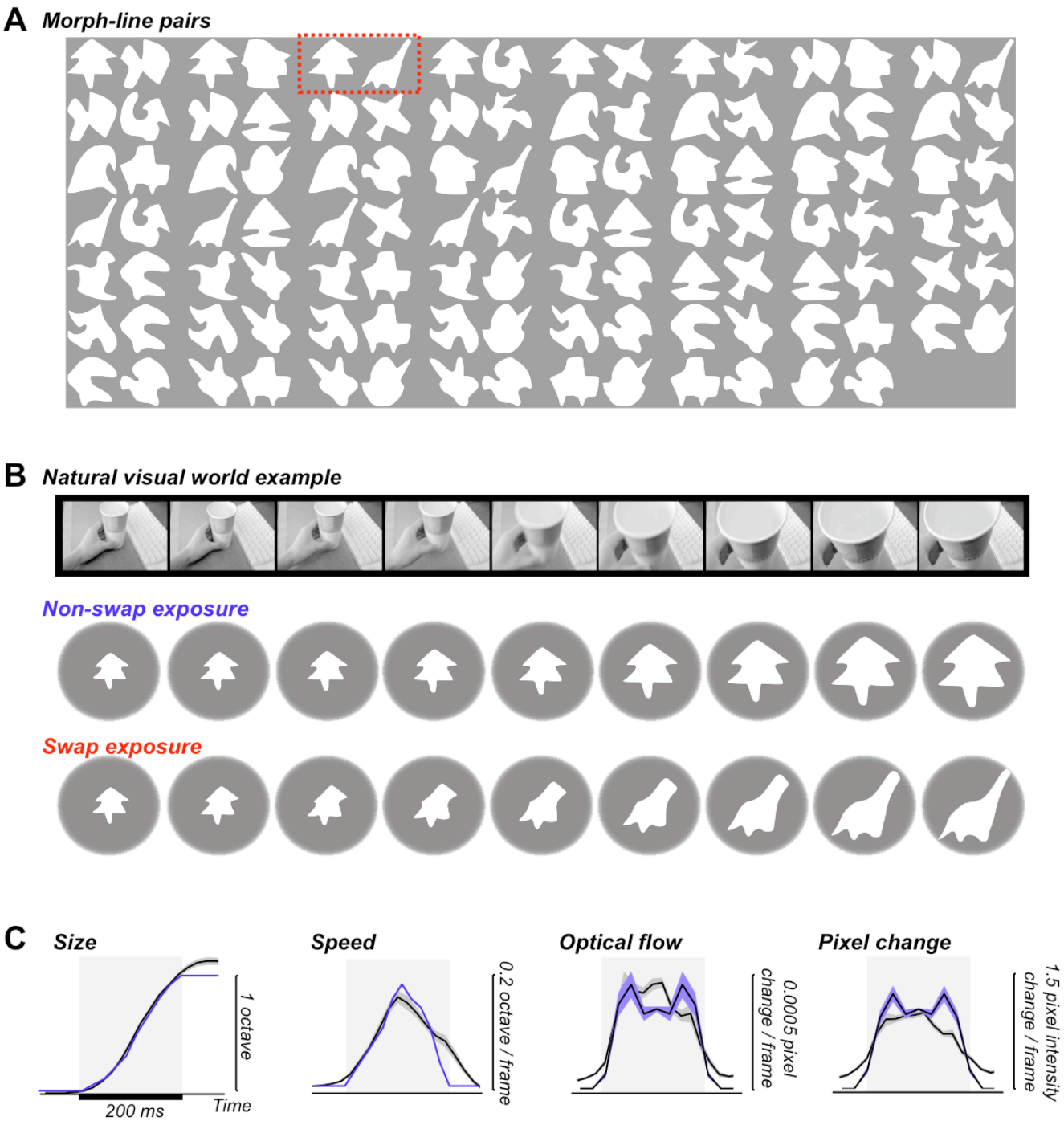


Figure S2. Stimuli from Experiment II and Comparisons to Natural Visual World Example.

(A) Cutout silhouette shapes were rendered using non-uniform rational B-spline. Each shape was rendered from a set of 24 control points. Matching and interpolating between the control points allowed us to parametrically morph between different shapes. Morph-lines were only achievable on a subset of all possible shape pairs in the silhouette class (Figure S1). The figure shows all the morph-line pairs used in Experiment II. Only Monkey 2 was tested in Experiment II given the stimulus class assignment. The example pair in (B) is highlighted.

(B) Top, a real world example of natural visual experience when lifting a cup to drink. Bottom, example exposure events we used in Experiment II (top, non-swap exposure event; bottom, swap exposure event). During each exposure event, the object size change was played out smoothly over a time period of 200 ms (frame rate: 40 frames/sec). We used the same dynamic (i.e. same size change profile but scaled in amplitude) for the two different types of size increase exposure events (1.5º→4.5º, 4.5º→9º, Figure 1B). For the object size decrease exposure events (4.5º→1.5º, 9º→4.5º, Figure 1B), the reverse sequence was played, which also mimicked the natural visual experience of putting down a cup (not shown).

(C) We quantified the statistics of the visual world example and our movie stimuli by a number of different image measures. Black lines show the visual world example (mean computed from videos of multiple repeats of the same action); blue lines show our movie stimuli (mean computed from all exposure events); shaded areas show SEMs. Object size was measured by the radius of the smallest bounding square around the shape (reported in units of octave, normalized to the initial size). Object size change speed was computed by taking the derivative of the object size measurements. Optical flow was computed using standard computer vision algorithm (Horn, 1986). Brightness patterns in the image move as the objects that give rise to them move. Optical flow quantifies the apparent motion of the brightness pattern. Here, mean optical flow magnitude over the entire image was computed. Pixel change was computed by taking the pixel intensity differences between adjacent video frames and the Euclidean norm of the pixel difference over the entire image was computed. All video frames were pre-processed to have unit variance before image analyses.

Supplemental Figure S3

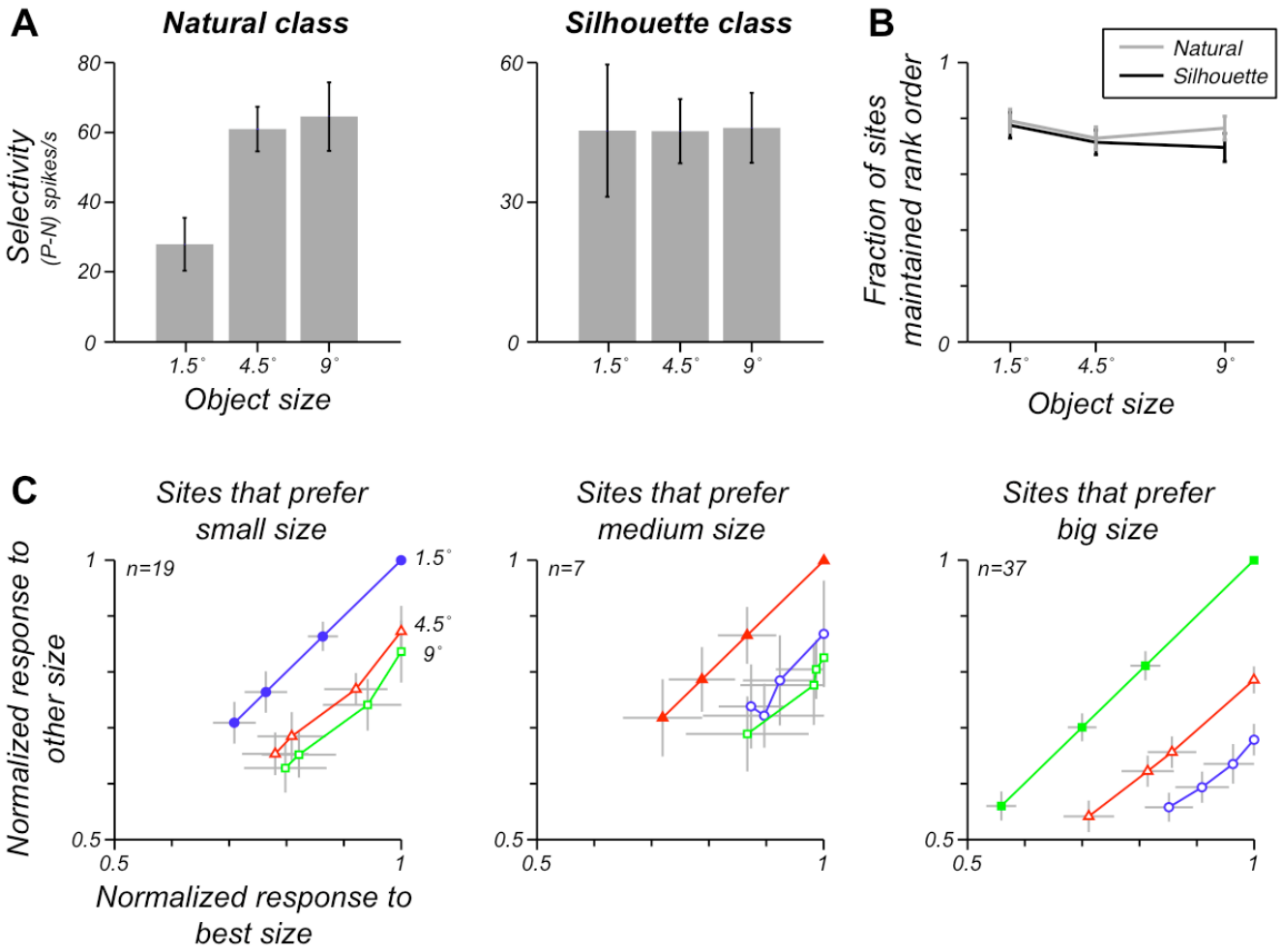


Figure S3. IT Multi-unit Activity Exhibits Size Tolerant Object Selectivity.

(A) IT neurons have object rank order selectivity that is largely unaffected by object size changes (Brincat and Connor, 2004; Ito et al., 1995; Logothetis and Sheinberg, 1996; Vogels and Orban, 1996), and that size tolerance is reflected in the IT multi-unit activity (Hung et al., 2005; Kreiman et al., 2006). Consistent with previous reports, most of the IT sites we recorded maintained their object rank order preference across the range of object size tested here (1.5º~9º). To quantify the degree of IT size tolerance for the swap and control object pairs, for each IT site we determined its preferred (P) and less preferred (N) object within an object pair using a portion of the response data at the medium object size (4.5º). We then used those "P" "N" labels to compute the object selectivity (P-N) from the remaining response data and for other object size. The plots show the mean ± SEM selectivity of all object selective sites from Experiment I and II (n=63). Positive selectivity indicates that IT sites, on average, maintained their object preference across size changes.

(B) Most of the individual IT sites (~80%, n=63) maintained their object rank order preference. The plot

shows the fraction of the IT sites in (A) that maintained their object rank order preference at each object size. Errorbars show SEMs.

(C) To summarize the average effect of object size changes on IT object selectivity across all four objects (swap and control object pairs combined), we split the 63 object selective IT sites into three groups based on their size preference. Preferred size for an IT site was defined as the size at which any object evoked the maximum response from the site. We then ranked the object preference based on the response at the preferred size (from best to worst). The abscissa represents the normalized response to the best object at each particular size. The ordinate represents the normalized response to the best object at the preferred size. Each data point shows the mean ± SEM. On average, IT sites maintained their object rank order preference. We found more sites preferring the extremity object sizes (1.5° and 9°) than the medium object size (4.5°), with more sites preferring the big object size (9°).

Supplemental Figure S4

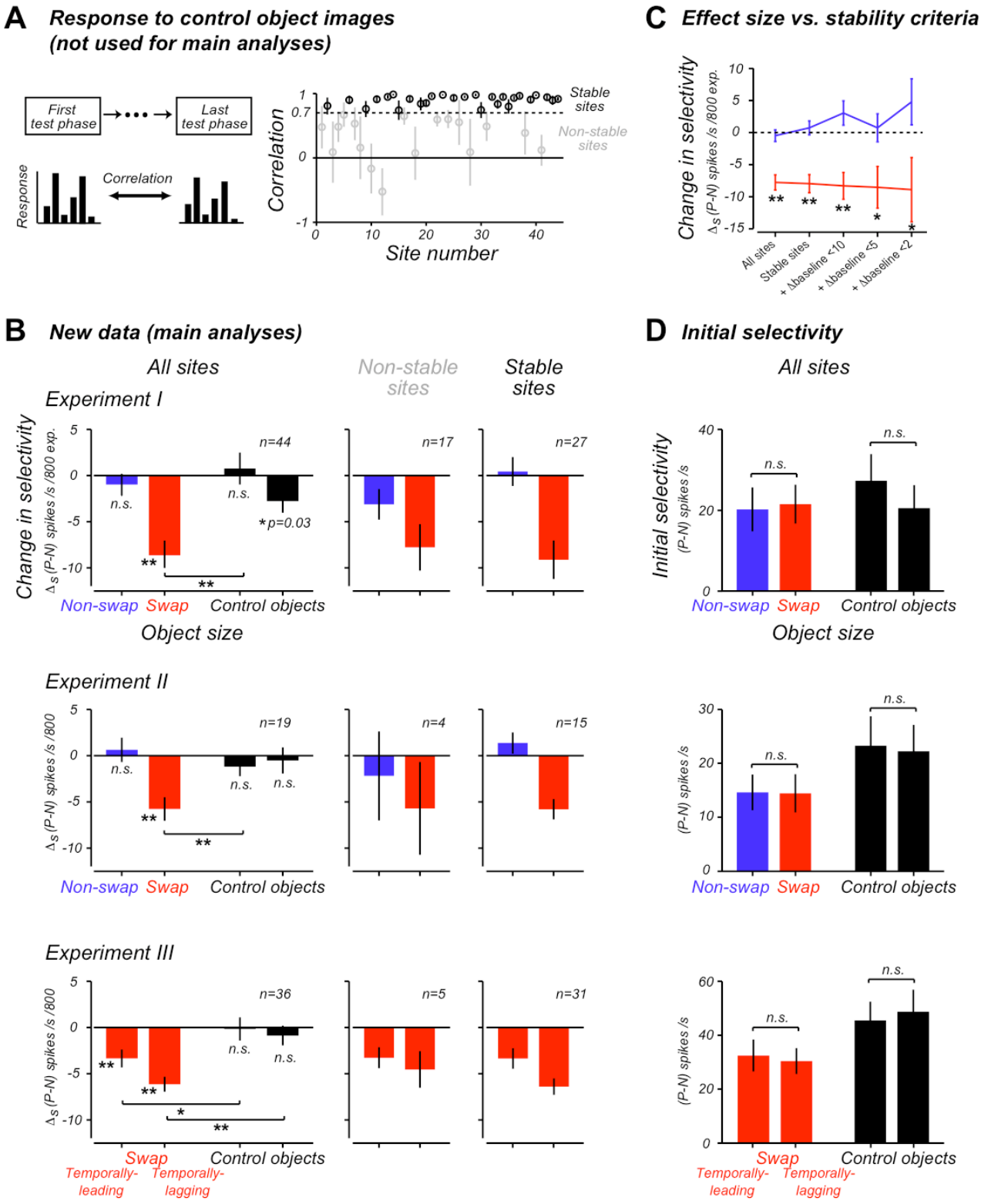


Figure S4. IT Results with All Object Selective Sites before and after Stability Screen.

(A) We deployed the key experience manipulation with a pair of swap objects (P and N) in the *Exposure Phase*. We also measured the IT response to a second pair of control objects (P' and N') along with the swap objects in the *Test Phase* (see Supplemental Experimental Procedures). We were interested in specific selectivity change in IT induced by our experience manipulation. However, there were potential non-specific changes in selectivity (e.g. from electrode drifts in tissue or tissue death) that could contaminate our effect of interest. Unlike traditional single-unit recording where one could judge the stability of long-term recording based on spike waveform, we did not have such measure in multi-unit recording. Thus we sought another independent measure of long-term recording stability (2-3 hours). To do this, we relied on IT selectivity among the images of the control objects (P' and N'). We picked these control objects to be sufficiently different from the swap objects in their pixel-wise similarity (Figure S1). Our analyses (panel B left column) and our previous investigation (Li and DiCarlo, 2008) have revealed that any experience-induced change in selectivity was specific to the swap objects. Leveraging this, we made the assumption that the control objects were far apart from the swap objects in IT shape space, thus they should be little affected by our experience manipulation. For each IT site, we computed Pearson's correlation between its response vectors to these control object images (6 dimensional vector, 2 objects x 3 sizes) measured from the first and last *Test Phase* (right panel: mean ± SEM; data from Experiment I only). A fraction of the sites showed low correlations, meaning their responses to the control object images had deviated from those measured in the first *Test Phase*. Note that a site could also have low correlation from having no tuning among the control object images to begin with, in those cases, we had no power to judge recording stability. In practice, we deemed a site stable if it had a correlation value higher than 0.7.

(B) All the main text results concentrated on the stable IT sites. Here, we present the main IT results from all object selective sites. Left column panels show mean ± SEM selectivity change, Δs(P-N), of the swap objects (red, swap size; blue, non-swap size) and control objects (black, same size as the swap objects). We found the change in IT selectivity was specific to the swap objects at the swap size. Statistically, object specificity of the selectivity change at the swap size was confirmed by a significant "object x exposure" interaction (p=0.009, repeated measures ANOVA). Next, we applied the stability screen outlined in (A) using the IT responses to the control object images (not used for the main analyses), we then looked to the change in selectivity, Δs(P-N), among the swap objects at the swap and non-swap size. The stability screen revealed non-specific changes in selectivity of the non-stable IT sites (middle column panels). Among the sites we deemed stable (right column panels), our experience manipulation induced very specific change in selectivity only at the swap size. Δs(P-N) was normalized to show IT selectivity change per 800 exposure events. * $p<0.05$ by t-test; ** $p<0.01$; n.s. $p>0.05$

(C) We also tested more strict forms of stability criteria that included baseline response (change <10, <5, and <2 spikes/s, before vs. after exposure) in addition to the standard stability screen. The plot shows Δs(P-N) at the swap (red) and non-swap size (blue). Data from Experiment I and II are combined (left to right: n=63; n=42; n=18; n=11; n=5). * $p<0.05$; ** $p<0.01$, t-test, swap vs. non-swap size.

(D) Mean ± SEM initial selectivity, (P-N), measured from the first *Test Phase*.

Supplemental Figure S5

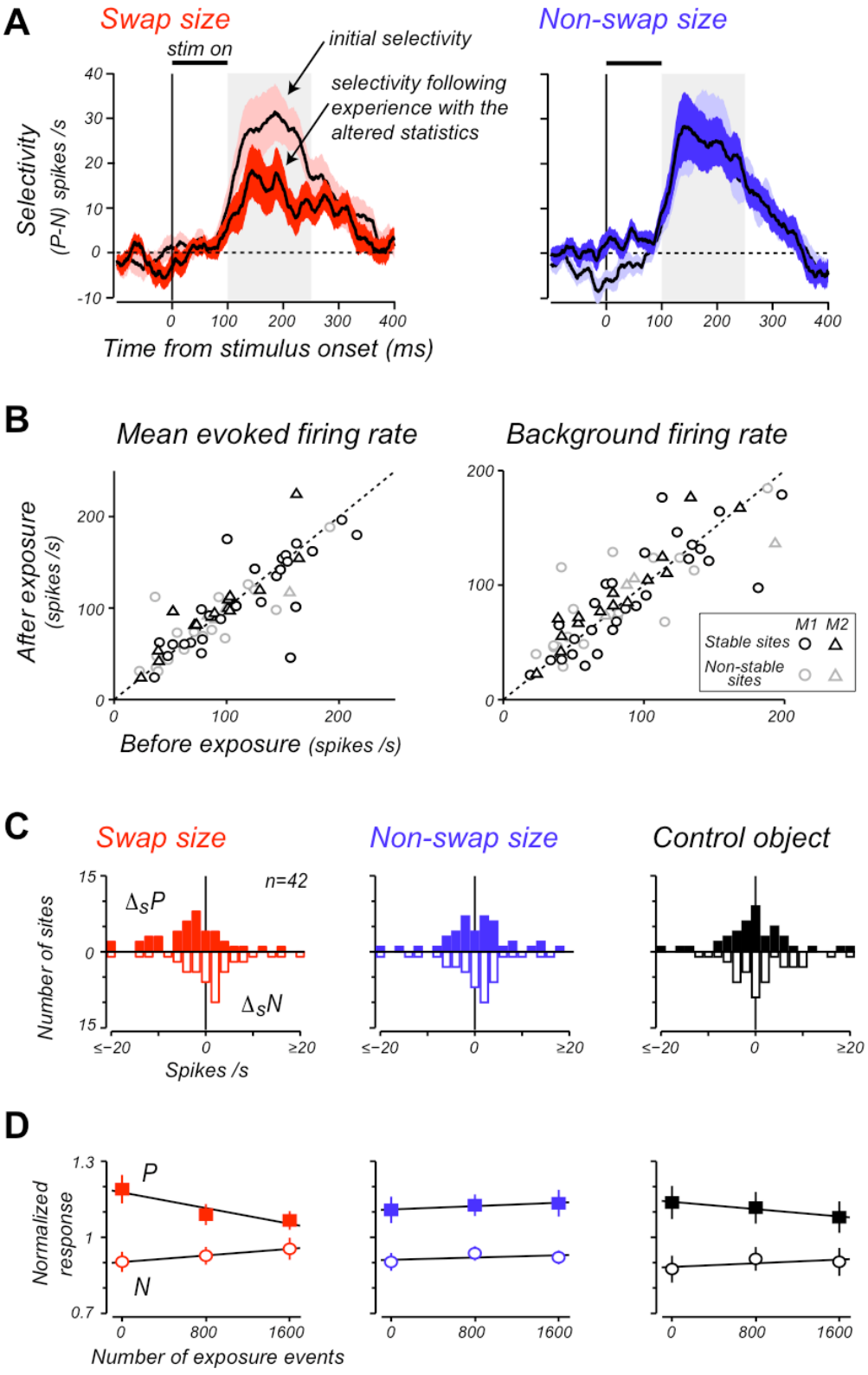


Figure S5. IT Response Changes Induced by Visual Experience.

(A) Mean ± SEM IT selectivity time course at the swap (left) and non-swap size (right) measured in the first (light colored) and last *Test Phase* (dark colored). Data from Experiment I and II are combined (n=42 IT sites). Gray region shows the standard spike count time window we used for all other analyses in the main text.

(B) IT firing rate was not altered by visual experience. For each IT site, we computed its mean evoked firing rate to all object images from the first and last *Test Phase*. All object selective sites were combined from Experiment I and II (n=63). We observed no net change in the mean evoked firing rate before and after our experience manipulation (left panel; p=0.24, two tailed t-test, before versus after). We also observed no net change in IT background firing rate (right panel; p=0.17, two tailed t-test). Background firing was measured from randomly interleaved blank stimulus presentations during the *Test Phases*. A few sites showed large change in their background firing rate even though they were classified as "stable sites" by their selectivity for the control object images (Figure S4). We thus tested more strict forms of stability criteria that included background firing rate with key results unchanged (Figure S4).

(C) We fit standard linear regression to each IT site's responses to object P and N at each object size as a function of the number of exposure events. The slope of the line fits (Δs) provided a measure of the response changes to P and N for each IT site. The histograms show the slope values of all the stable sites from Experiment I and II (n=42). ΔsP and ΔsN were normalized to show response changes per 800 exposure events.

(D) Mean ± SEM normalized responses to object P and N as a function of the number of exposure events. For each IT site, response of each *Test Phase* was normalized to the mean response to all object images in that *Test Phase*.

Supplemental Figure S6

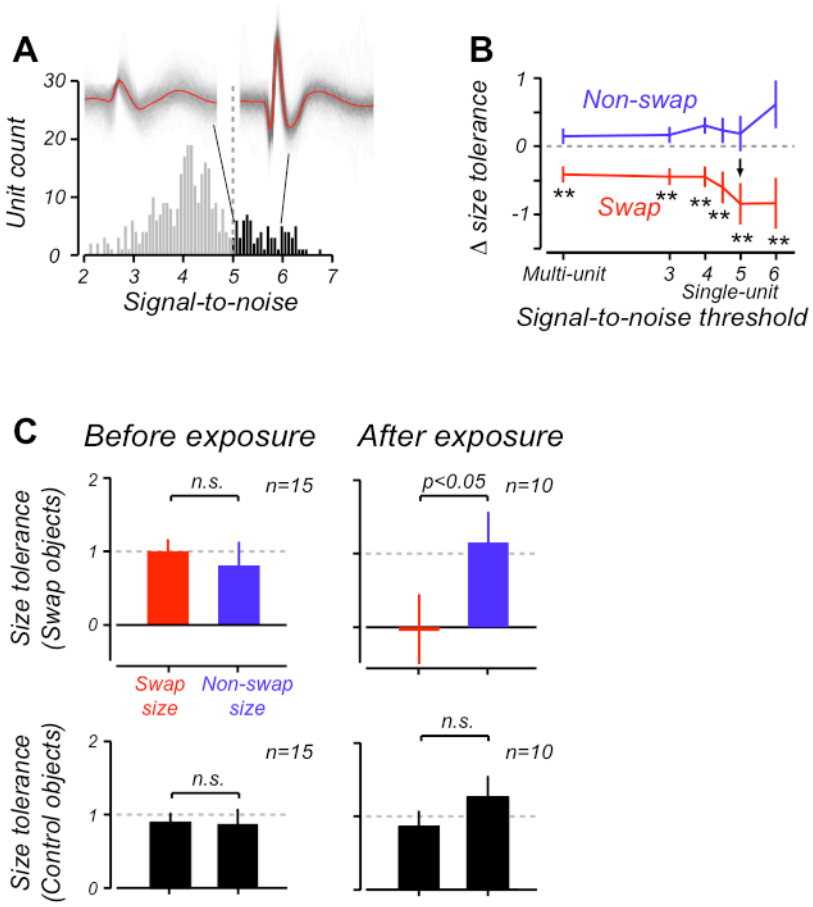


Figure S6. IT Single-Unit result is Robust to Unit Selection Criteria.

(A) We performed PCA-based spike sorting on the waveforms collected during each *Test Phase*, treating each unit as an independent sample from the IT population either before or after the altered visual experience. Each unit obtained from the spike sorting was further evaluated by its signal-to-noise ratio (SNR: ratio of peak-to-peak mean waveform amplitude to standard deviation of the noise). The histogram shows the distribution of SNR for all the units obtained. For all the single-unit analyses in the main text (Figure 4), we set a SNR threshold (dash-line: SNR=5.0) above which we will term a unit "single-unit".

(B) To ask if the result was robust to our choice of the single-unit SNR threshold, we systematically varied the threshold and re-performed the same analyses. The plot shows the experience-induced change in size tolerance (Δ size tolerance, same as in Figure 4D) at the swap (red) and non-swap (blue) size. We found that the result was highly robust to the single-unit selection criteria, and the experience induced effect at the swap size only grew stronger when we increased the strictness of the single-units criteria. ** p<0.001, bootstrap; arrow head shows the single-unit threshold used in Figure 4.

(C) Mean ± SEM size tolerance for the swap and control objects measured in the same population of neurons. Same as Figure 4B.

Supplemental Figure S7

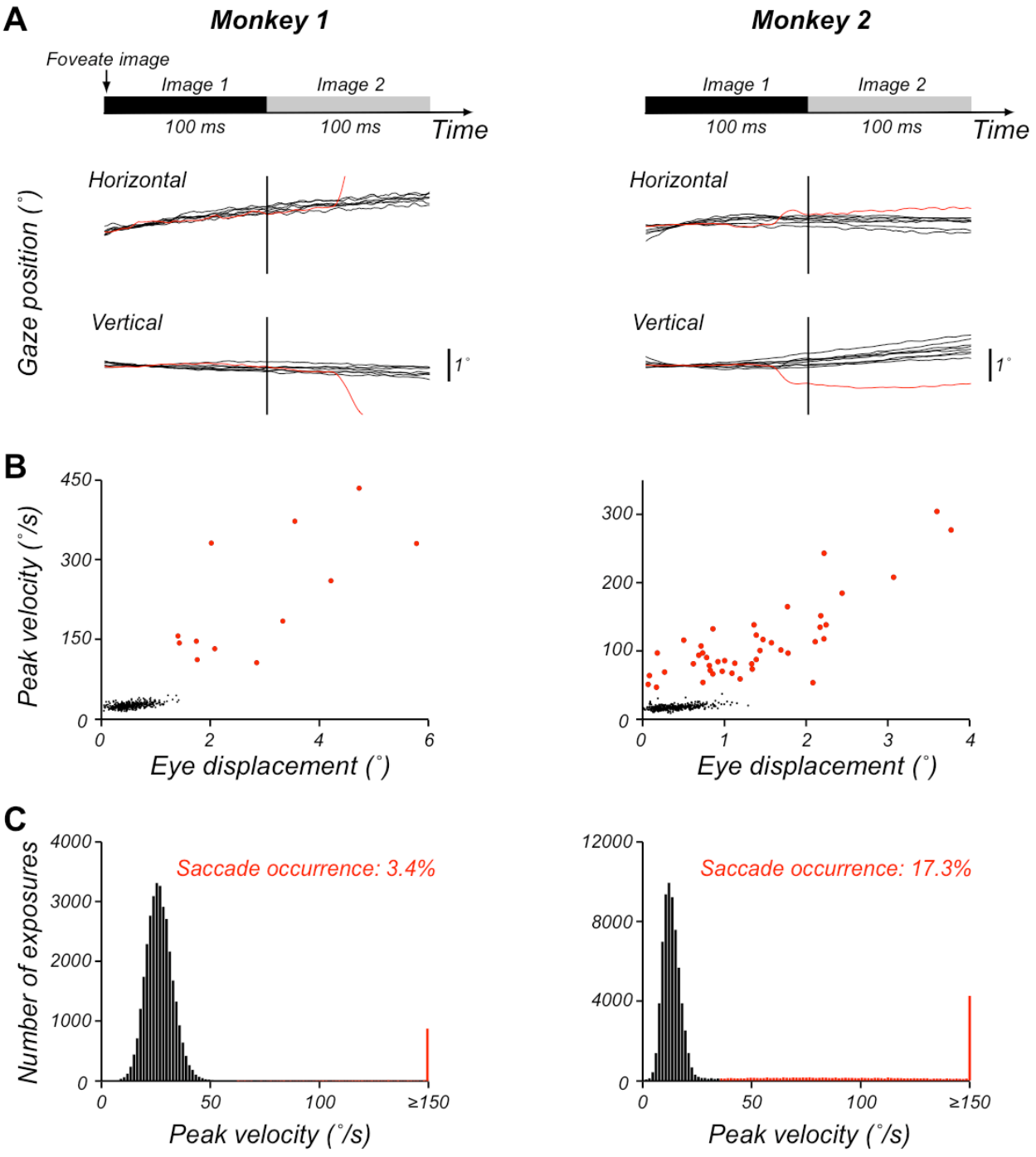


Figure S7. Eye Movement Pattern during Exposure Events.

(A) Our previous study on IT position tolerance learning (Li and DiCarlo, 2008) showed that unsupervised experience of temporally contiguous images coupled by an intervening saccade can reshape IT position tolerance. Here, we showed that unsupervised experience of temporally contiguous images presented on animals' center of gaze is sufficient to induce IT size tolerance learning. The animals freely viewed a gray computer screen on which objects intermittently appeared at random position. We deployed the experience manipulation (i.e. image pairing across time) during brief periods of the animals' fixation. The exposure events were meant to mimic regimes of natural vision where object change size on the retinal due to object motion (de-coupled from intervening eye movements). However, it was possible that the discontinuous image changes we employed always induced small saccades from the animals during the exposure events, hence the observed IT size tolerance learning is simply the same piece of phenomenology as the IT position tolerance learning reported before. Here we examine this possibility by analyzing the *Exposure Phase* eye movement data around the time of image change (±100ms). The plots show the stimulus presentation time sequence (top) and aligned eye position data (bottom) during a few exposure events from one example *Exposure Phase*. The animals were able to maintain their gaze position throughout the periods of image change in most cases, though there were minor drifts (typically <1º). Occasionally, the animals made small saccades (red eye traces), however, these only constituted a small fraction of all exposure events, see (C).

(B) All the eye movement data from the example *Exposure Phase* was plotted in their relationship between the total eye displacement and peak velocity around the time of image change (±100ms). Each data point represents data from one exposure event (i.e. one trace in (A)). For saccades (red dots), there was a systematic relationship between the peak velocity and eye displacement (i.e. main sequence), which distinguished itself from the pattern of fixation eye movement (black dots). There was always good separation between the two types of eye movement pattern, thus we used a peak velocity threshold to define saccades (Monkey 1: ~60º/s; Monkey 2: ~40º/s).

(C) Histograms of eye movement peak velocity during all exposure events (Experiment I population data: all *Exposure Phases* across all recording sessions were combined). Exposure events that contained saccades are shown in red bins and exposure events without saccades are in black bins. The animals made saccades only on a small fraction of all exposure events (Monkey 2 was slightly worse). Given the small occurrence of saccades in comparison to our previous study on position tolerance where saccades accompanied every exposure event (Li and DiCarlo, 2008), we concluded that the possibility of intervening saccades cannot account for the observed IT selectivity change.

Supplemental Figure S8

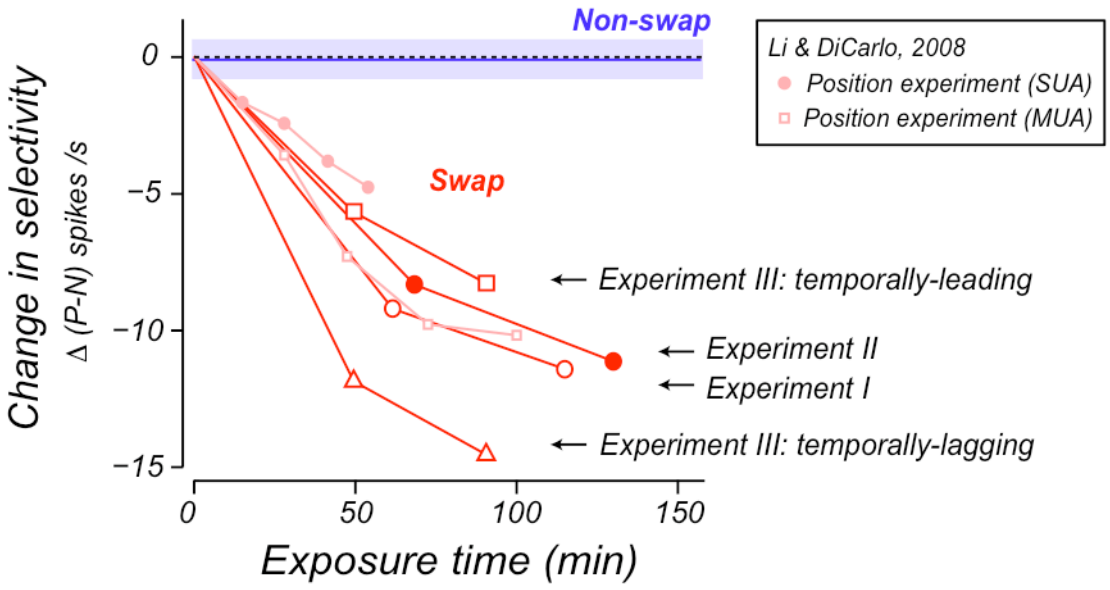


Figure S8. Effect Size Comparisons across Different Experience Manipulations as a Function of Exposure Time.

Mean change in IT object selectivity, Δ(P-N), as a function of swap exposure time for different experience manipulations (i.e. Experiments I, II, III; position experiments: Li and DiCarlo, 2008). Exposure time was determined based on the time *Test Phase* data files were saved. For each data points, we computed the average exposure time across all the neurons/sites (grouped by their *Test Phase* numbers). Plot format is the same as main text Figure 5. Mean ± SEM selectivity change at the non-swap size (or position) is shown in blue (pooled across all experiments). SUA: single-unit activity; MUA: multi-unit activity.

Supplemental Figure S9

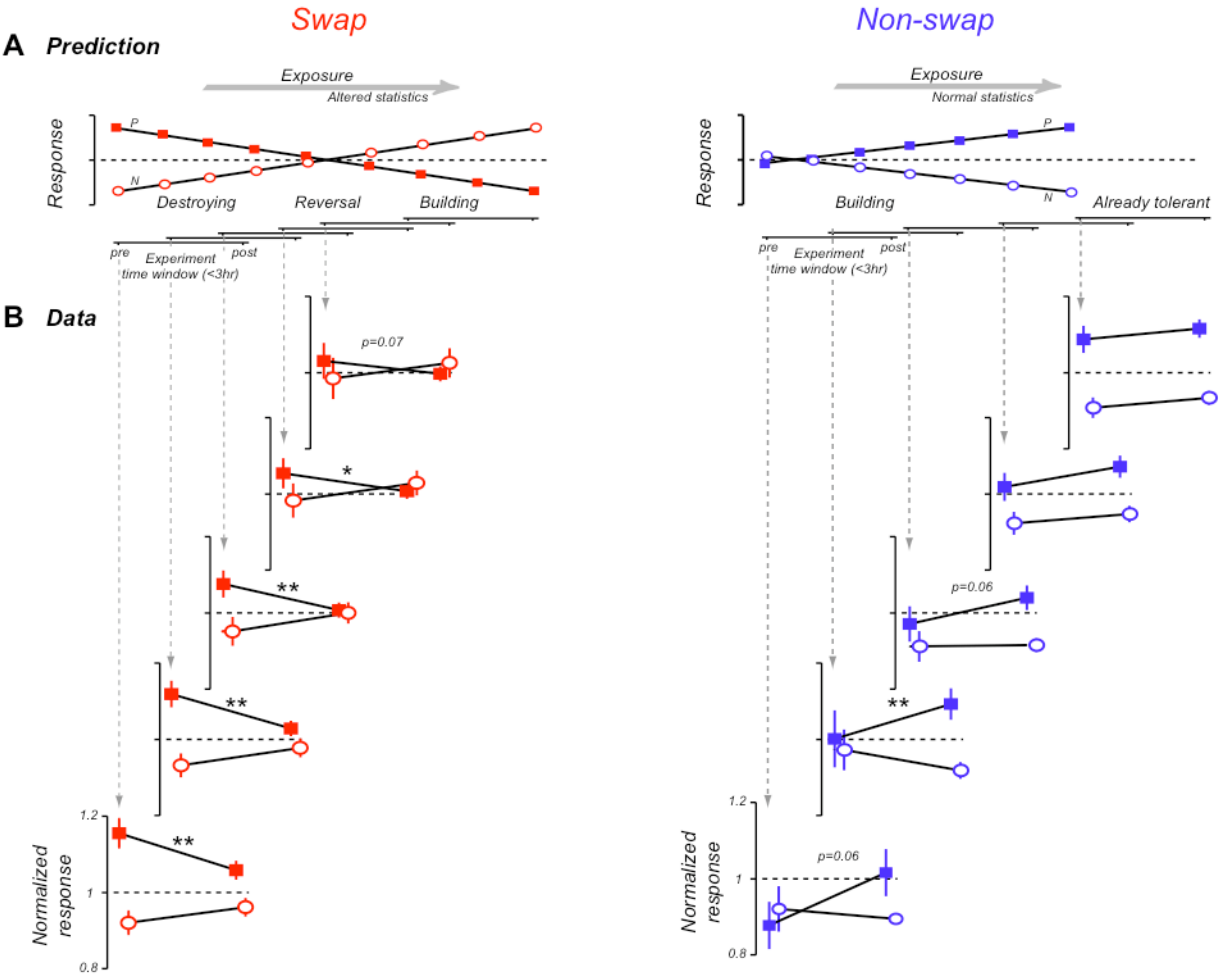


Figure S9. Breaking and Building Tolerant selectivity, Size Experiment Data.

Mean ± SEM normalized response to object P and N at the swap size (A) and non-swap size (B) among sub-populations of IT multi-unit sites. Other details same as Figure 6 and 7. Size experiment data only (Experiment I and II).

Supplemental Experimental Procedures

Visual Stimuli

Stimuli were presented on a 21" CRT monitor (85 Hz refresh rate, ~48 cm away, background gray luminance: 22 Cd/m2, max white: 46 Cd/m^2). We used 96 achromatic images from two classes of visual stimuli: 48 cutout natural objects and 48 silhouette shapes, both presented on gray background (Figure S1). We chose these two classes of stimuli to be sufficiently different from each other in their pixel-wise similarity (Figure S1), so that neuronal plasticity induced among one object class would be unlikely "spill-over" to the other class (our results and previous work confirmed this assumption, see Figure S4 and Li and DiCarlo, 2008). All stimuli were presented on the animal's center of gaze during IT selectivity testing. In all experiments, we always used three object sizes (1.5º, 4.5º, 9º, in Experiment I and II; 1.5º, 3º, 4.5º in Experiment III). Object size was defined as the width of the smallest bounding square to contain the object. The medium object sizes were used to pick preferred (P) and non-preferred (N) objects for an IT site in an initial screening (see *Neuronal Assays* below), but we designed our manipulations and analyses to focus on the two extremity sizes (Figures 1B, 1C, 8A).

In Experiment II, to create the smoothly-varying identity-changing movie stimuli, we created morph lines between a subset of the silhouette shapes. Seven intermediate morphs were created in-between each object pairs. The movie stimuli were created to match the dynamics of object size changes that could be encountered in the natural world (see Figure S2).

Behavioral Assay

Custom software controlled the stimulus presentation and behavioral monitoring. Eye position was monitored in nearly real-time (lag of ~3 ms) using standard sclera coil technique (Robinson, 1963) and in-house software, and saccades >0.2° were reliably detected (DiCarlo and Maunsell, 2000).

Test Phase: During each *Test Phase* (~10 minutes), IT neuronal selectivity was probed in two different tasks. Monkey 1 freely searched an array of eight small dots (size 0.2º) vertically arranged 3º apart. The dots never changed in appearance, but on each "trial", one dot would be randomly baited in that a juice reward was given when the animal foveated that dot, and the next "trial" continued uninterrupted. Typically, the monkey saccaded from one dot to another (not always the closest dot) looking for the hidden reward. During this task, object images were presented (100 ms duration) on the animal's center of gaze, (onset time was the detected end of a saccade; approximately one such presentation every other saccade, never back-to-back saccades). Thus, the monkey's task was unrelated to these test stimuli. To limit unwanted experience with the visual stimuli, each such presented object was immediately removed upon detection of any saccade and these aborted presentations were not included in the offline analyses. Monkey 2 performed a more standard fixation task in which it foveated a single, central dot (size 0.2º, ±1.5º fixation window) while object images were presented at a

natural, rapid rate (5 images/s; 100 ms duration, 100 ms blank intervals). Reward was given at the end of the trial (5-8 images presented per trial). Upon any break in fixation, any currently present object image was immediately removed (and not included in the analyses), and the trial aborted. The animal could typically maintain fixation successfully in >75% of the trials. Aside from the task differences (free-viewing search vs. fixation), retinal stimulation in the two tasks was essentially identical. ~60 (±2) repetitions of each image were collected in the first *Test Phase* and ~50 (±2) repetitions in all the later *Test Phases*.

Exposure Phase: During each *Exposure Phase* (~1.5 hr), the animal freely viewed the monitor while object images (pseudo-randomly chosen) intermittently appeared at random positions on the screen. Because foveating a suddenly appearing object is a natural, automatic behavior, essentially no training was required, and the monkey almost always looked directly to the object (>90% of the time). 100 ms after the animal had foveated the object (defined by a saccade offset criteria of eye velocity<10º/s and a ±1.5º window centered on the object), the object underwent a size change on the animal's center of gaze. Importantly, some of the object size changes were accompanied by identity changes (i.e. our key manipulation, see details of specific experiments in the main text Experimental Procedures). The free viewing was meant to keep the monkey engaged in natural visual exploration, but the manipulation of object size statistics was always deployed during the brief intervals of fixation during natural exploration (see eye movement analyses in Figure S7). The animal was only rewarded for looking to the object to encourage exploration, thus no explicit supervision was involved. There were a total of 8 different exposure event types in the full design (illustrated by the eight arrows in Figure 1B). One *Exposure Phase* consisted of 1600 exposure events: 200 exposure events per arrow exactly.

Neuronal Assay

Muti-unit activity (MUA) was gathered from 154 IT sites (n=44 for Experiment I; 19 for Experiment II; 91 for Experiment III) by randomly sampling over a ~4x4 mm area of the ventral STS and ventral surface lateral to the AMTS (Horsey-Clark coordinates: AP 13-17 mm; ML 18-22 mm at recording depth) from the left hemispheres of two monkeys. MUA was defined as all the signal waveforms in the spiking band (300 Hz – 7 kHz) that crossed a threshold set to ~2 s.d. of the background activity. That threshold was held constant for the entire session. A snippet of waveform data sampled at 0.07 ms intervals was recorded for 8 ms around each threshold-triggering event and saved for offline spike sorting (see *Data Analyses* below).

Each day, a glass shielded platinum-iridium microelectrode wire was introduced into the brain *via* a guide-tube and advanced to the ventral surface of the temporal lobe by a hydraulic microdrive (guided by anatomical MRI). We then advanced the microelectrode while the 96 object images (Figure S1) were pseudo-randomly presented on the animals' center of gaze (animal tasks identical to those in the *Test Phases*). Once a visually driven recording site was found (based on online inspection), we stopped advancing and left the electrode in the brain to allow for tissue settling (up to 2 hours) before the recording session started. Each recording session began with an initial screening in which the IT sites were probed with the same object set (96 objects, ~10 repetitions per object, all presented on the center of gaze) for object pair

selection:

Main Test Objects (Swap Objects): Among the objects that drove the site significantly above its background response (t-test against randomly interleaved blank presentation, $p<0.05$, not corrected for multiple tests), the most preferred (P) and least preferred (N) objects were chosen as a pair. Thus, both objects tended to drive the neuronal recording site, and most sites had selectivity for one (P) over the other (N). These two objects were chosen subject to the condition that both objects were from the natural object class (Monkey 1) or both were from the silhouette object class (Monkey 2; see Figure S1).

Control Objects: For each recorded IT site, we also used the same initial screening (above) to choose a second pair of control objects (P' and N'). Our goal was to choose two objects the IT site was selective for but were very distant from the swap objects in IT shape space. Because we do not know the dimensions of IT shape space, we cannot strictly enforce this. In practice, we simply ensured that the control objects were always chosen from the object class that was not used for the swap objects (i.e. the silhouette object class for Monkey 1, and the natural object class for Monkey 2, see Figure S1). Within this constraint, the control objects were chosen using the exact same responsivity and selectivity criteria as the swap objects (described above).

Once the initial screening and object selection was completed, we then carried out the *Test* and *Exposure Phases* in alternation while making continuous recording from the IT site for the entire recording session (~3 hours). The swap objects and control objects were each tested at all three sizes in each *Test Phase* but only the swap objects were shown and manipulated during the *Exposure Phases*.

Data Analyses

Neuronal data recorded from the 154 IT sites was first tested for their object selectivity. Offline analyses revealed that a fraction of the sites were not significantly selective among the swap object pairs (two-way ANOVA, 2 object x 3 sizes, $p>0.05$ for both "object" main effect and "object x size" interaction), probably because only a limited number of response repetitions were collected during the initial screening and we selected the objects to both produce a statistically significant response (as described above). We excluded those sites and only concentrated on the remaining object-selective sites (n=43 for Experiment I; 19 for Experiment II; 36 for Experiment III, many sites from Experiment III showed significant selectivity only for the swap object pair or only for the control object pair, but we concentrated on the sites that showed significant selectivity for both the swap and control object pairs). These sites were subject to one more screening for recording stability (see below) and all the results presented in the main text were from the object-selective and stable sites (n=27 for Experiment I; 15 for Experiment II; 31 for Experiment III).

Recording Stability Screen: We were interested in specific selectivity changes induced by our experience manipulation. However, we were concerned that non-specific selectivity changes (e.g. resulting from electrode drifts in tissue or neuronal injury) could potentially contaminate our effect of interest. Our controls were designed to make sure that we would not interpret any such effects as evidence of learning, but we still wanted to do our best to insure

that any non-specific effects would not mask the size of our effect of interest. Unlike single-unit recording where one can judge the stability of recording based on spike waveform isolation, we do not have such measures in multi-unit recording. Thus we sought another independent measure of recording stability. To do this, we relied on the selectivity among the control object images (see above). We proceeded under the assumption that these control object images were far apart from the swap object pairs in the IT shape space, there should be little change in the selectivity among these control object images induced by our experience manipulation (our results and previous work confirmed this assumption; see Supplemental Figure S4 and Li and DiCarlo, 2008). That is, the response to these objects provides a gauge of any non-specific changes in IT selectivity. To quantify that gauge, we computed Pearson's correlation between the control image response vectors (6 dimensional vector, 2 objects x 3 sizes) measured from the first and last *Test Phases*. We deemed an IT site "stable" if it had a correlation value higher than 0.7 (Figure S4). In the main text, we only present results from these stable sites because they provide the cleanest look at our data and the best quantitative measure of learning magnitude. Critically, this site selection procedure relies only on data that is fully independent of our key exposure condition and key control condition (e.g. Figure 1B), so there is no selection bias. Nevertheless, we also repeated the same analyses on all of the recorded IT sites and found that the main results were qualitatively unchanged (see Figure S4). We also tested more strict forms of stability criteria that included background activity change (<10, <5, and <2 spikes/s). With these stability criteria, all the key results also remained the same (Figure S4C).

Computing (P-N) neuronal selectivity: To avoid any bias in this estimate of selectivity, for each IT site, we set aside an independent set of response data from the first *Test Phase* (10 response repetitions to each object in each size) and used those data only to define the labels "P" and "N" ("P" was taken as the object that elicited a bigger overall response pooled across object size). We recorded 10 extra response repetitions in the first *Test Phase* in anticipation of this need for independent data (60 repetitions in the first *Test Phase*, 50 repetitions in the later *Test Phases*). The label "P" and "N" for the site was then held fixed across object size and later *Test Phases*, and all remaining data was used to compute the selectivity (P-N) using these labels. This procedure ensured that any observed response difference between object P and N reflected true selectivity, not selection bias. Because different splitting of screen and remaining data may not result in consistent "P" "N" label, for each IT site this procedure was performed 100 times (different splitting of screen and remaining data in the first *Test Phase*) to obtain an averaged selectivity estimate (P-N). Variability arising from this procedure is reflected in the error bars of Figure 2C and 3B for each IT site.

Statistical Tests for the "Size x Exposure" Interaction: The key part of our experimental prediction is that any change in object selectivity should be found predominantly at the swap size (Figure 1C). To directly test for such an interaction between object size and our independent variable (exposure), we performed two different statistical tests on the neuronal selectivity measurements (P-N, in units of spikes/s). This main prediction and statistical results are from pooling across neurons (i.e. pooled "subjects" design with counterbalance).

First, we applied a two-factor repeated measures ANOVA. To design the test, we treated each IT site as one repeated measurement (i.e. one subject) with two within-group factors

("exposure" and "size"). Repeated measures ANOVA expects that all subjects are measured across the same number of conditions, however, our data was such that each IT site was tested for differential amount of time: some IT sites had three *Test Phases* while others only had two (due to different rates of experimental progress on each day and normal variation in the animal's daily work ethic). To get around this problem, for each IT site, we simply used the data only from the first and last *Test Phase*, omitting the data from the intermediate *Test Phases* for some IT sites. Thus in our ANOVA design, the "exposure" factor had two levels, and the "size" factor also had two levels: swap and non-swap. Our main focus was on the significant interactions between "exposure" and "size" (see main text). Our data also revealed significant main effects of "exposure" (Experiment I: p=0.0004; Experiment II: p=0.014) and no significant main effect of "size" (p = 0.72; p = 0.32). Given our experience manipulation and counterbalanced experience design across object size, this pattern of main effects is expected under the temporal contiguity hypothesis (see Figure 1C).

We also carried out a second, more non-parametric statistical test for the interaction of "exposure" and "size" by applying a general linear model. The formulation is similar to ANOVA. However, it is not subject to assumptions about the form of the trial-by-trial response variability. We have previously used the same method in our study on IT position tolerance learning (Li and DiCarlo, 2008) and simulations with Poisson spiking neurons have confirmed the correctness of our analysis code (~5% significant occurrence at p<0.05 with null effects). The model had the following form:

$$(P-N)_{neuron=n,size=s,\exp osure=e} = a_n + b_1 \cdot s + b_2 \cdot e + b_3 \cdot (s \cdot e)$$

The three independent variables of the model were: "size" (*s*), "exposure" (*e*), and their interaction (i.e. their product, *s*·*e*). The "size" factor had two levels (i.e. *s* = 1 for swap size, -1 for non-swap size) the "exposure" factor had up to three levels depending how long a site was tested, (i.e. *e* = 0 for pre-exposure, and could be up to 1600 exposures in increments of 800's). Each a_n was the selectivity offset specific to each IT site; b_1, b_2, and b_3 were slope parameters that were shared among all the sites (i.e. within subject factors). Thus, the complete model for our population of *n* sites (*n*=27, Experiment I; n=15, Experiment II) contained a total of *n*+3 parameters that were fit simultaneously to our entire data set. The a_n's absorbed the site-by-site selectivity differences that were not of interest here, and the remaining three parameters described the main effects in the population, with b_3 of primary interest (interaction).

We fit the linear model to the data (standard least squares), and then asked if the observed value of the interaction parameter (b_3) was statistically different from 0. To do this, we obtained the variation of the b_3 estimate *via* bootstrap over both IT sites and repetitions of each site's response data. The exact procedure was done as follows: for each round of bootstrap over IT sites, we randomly selected (with replacement) *n* sites from our recorded *n* sites, so a site could potentially enter one round of bootstrap multiple times. Once the sites were selected, we then randomly selected (with replacement) the response repetitions included for each site (our unit of data here was a scalar spike rate in response to a single repetition of one object image in one size). Importantly, the selection of the response repetitions was done after we have excluded 10 response repetitions reserved for determining object labels ("P" and "N"). This absolute independence of the data allowed us to obtain unbiased selectivity estimates. Each site's (P-N) was computed from its selected response repetitions. The linear model was then

fit to the data at the end of these two random samples to obtain a new b_3 estimate. This procedure was repeated 1000 times yielding a distribution of b_3 estimates, and the final p-value was computed as the fraction of that distribution that was less than 0. This p-value was interpreted as: if we were to repeat this experiment, with both the variability observed in the neuronal responses as well as the variability in which IT sites were sampled, what is the chance that we would *not* see the interaction observed here? In effect, this bootstrap procedure allowed us to derive a confidence interval on the model parameter estimate (b_3), and the duality of confidence intervals and hypotheses testing allowed us to report that confidence interval as a p-value (Efron and Tibshirani, 2003).

Statistical Tests for the Response Change in Single Sites: We evaluated each IT multi-unit site's selectivity (P-N) change by fitting linear regression as a function of the number of exposure events to obtain a slope, Δs(P-N). The statistical significance of the response change for each IT site was evaluated by permutation test. Specifically, for each site, we randomly permuted the *Test Phase* label of the response data (i.e. which *Test Phase* each sample of P and N response data belonged to, our unit of data here was a scalar spike rate in response to a single repetition of one object image in one size). We then re-computed the (P-N) selectivity on the permuted data and fit the linear regression. The permutation procedure was performed 1000 times to yield a distribution of slopes (empirical "null distribution" of Δs(P-N)). The p-value was determined by counting the fraction of the null distribution that exceeded the linear regression slope obtained from the data. All sites with $p < 0.05$ were deemed significant (see main text).

Combining the Position and Size Tolerance Learning Data: In main text Figures 6 and 7, we pooled the data from size experiment I, II, (n=42 MUA sites), and our previous position tolerance experiment (n=10 MUA sites collected using the same method described above, see Li and DiCarlo, 2008) because the two experiments used similar experience manipulations and the effect magnitude was comparable (Figure 5). To enter this analysis, we required that the sites had (P-N) selectivity at the medium object size/position (>5 spikes/s and <50 spikes/s, n=34). This was done under the logic that such selectivity is needed to provide a driving force for learning. We then used independent data to divide the sites into different groups based on the selectivity at the swap position/size in Figure 6 (Group 1: all sites; Group 2: <40; Group 3: <20; Group 4: <10; Group 5: <5; Group 6: <0) or at the non-swap position/size in Figure 7 (Group 1: <0; Group 2: <5; Group 3: <10; Group 4: <20; Group 5: all sites). We used independent data to select these sub-populations so that any stochastic fluctuations in site-by-site selectivity would produce no average selectivity change.

Single-unit Sorting and Analyses: We performed principle component analyses (PCA) based spike sorting on the waveform data collected during each *Test Phase*. K-mean clustering was performed in the PCA feature space to yield multiple units. The number of clusters was determined automatically by maximizing the distances between points of different clusters. Each unit obtained from the clustering was further evaluated by its signal-to-noise ratio (SNR: ratio of peak-to-peak mean waveform amplitude to standard deviation of the noise). For the analyses presented in Figure 4, we set a SNR threshold of 5.0, above which we will term a unit "single-unit". We verified that the key result is robust to the choice of this threshold (Figure S6).

Because there was a great amount of cell-to-cell variability in IT neurons' selectivity, we computed a normalized selectivity measure for each neuron (Figure 4). Each neuron's size tolerance was computed as $(P-N)/(P-N)_{medium}$, where (P-N) is the selectivity among the two objects at the tested size and $(P-N)_{medium}$ is the selectivity at the medium object size. A size tolerance of 1.0 means that a neuron perfectly maintained its selectivity across the size variations spanned here. Because not all the single-units had object selectivity, only units that showed selectivity at the medium size were included ($(P-N)_{medium}$>1 spikes/s).

Supplemental References

Brincat, S.L., and Connor, C.E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat Neurosci *7*, 880-886.

DiCarlo, J.J., and Maunsell, J.H.R. (2000). Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. Nat Neurosci *3*, 814-821.

Efron, B., and Tibshirani, R.J. (2003). An Introduction to the Bootstrap. (Chapman & Hall).

Horn, B. (1986). Robot Vision (MIT Press).

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863-866.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. Journal of Neurophysiology *73*, 218-226.

Kreiman, G., Hung, C.P., Kraskov, A., Quiroga, R.Q., Poggio, T., and DiCarlo, J.J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron *49*, 433-445.

Li, N., and DiCarlo, J.J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science *321*, 1502-1507.

Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. Ann. Rev. Neurosci. *19*, 577-621.

Robinson, D.A. (1963). A method of measuring eye movements using a scleral search coil in a magnetic field. IEEE Transactions on Biomedical Engineering *101*, 131-145.

Vogels, R., and Orban, G.A. (1996). Coding of stimulus invariances by inferior temporal neurons. Prog Brain Res *112*, 195-211.