Inferotemporal cortex reflects behaviorally-relevant target match information during invariant object search

Noam Roth and Nicole C. Rust*

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104

* Corresponding author
Email: nrust@psych.upenn.edu

Short title: Inferotemporal cortex during invariant object search

Abstract:

Finding a sought visual target object requires combining visual information about a scene with a remembered representation of the target to create a “target match” signal that indicates when a target is in view. Target match signals have been reported to exist within high-level visual brain areas including inferotemporal cortex (IT), where they are mixed with representations of image and object identity. However, these signals are not well understood, particularly in the context of the real-world challenge that the objects we search for typically appear at different positions, sizes, and within different background contexts. To investigate these signals, we recorded neural responses in IT as two rhesus monkeys performed a delayed-match-to-sample object search task in which target objects could appear at a variety of identity-preserving transformations. Consistent with the existence of behaviorally-relevant target match signals in IT, we found that IT contained a linearly separable target match representation that reflected behavioral confusions on trials in which the monkeys made errors. Additionally, target match signals were highly distributed across the IT population and while a small fraction of units reflected target match signals as target match suppression, most target match signals were reflected as target match enhancement. Finally, we found that the IT target match representation was configured to minimally interfere with IT visual representations. Together, these results support the existence of a robust, behaviorally-relevant target match representation in IT that is multiplexed with IT visual representations.

Introduction:

Finding a sought visual target object requires combining incoming visual information about the identities of the objects in view with a remembered representation of a sought target object to create a “target match” signal that indicates when a target has been found. During visual target search, target match signals have been reported to emerge in the brain as early as visual areas V4 (Bichot et al., 2005; Chelazzi et al., 2001; Haenny et al., 1988; Kosai et al., 2014; Maunsell et al., 1991) and IT (Chelazzi et al., 1998; Chelazzi et al., 1993; Eskandar et al., 1992; Gibson and Maunsell, 1997; Leuschow et al., 1994; Mruczek and Sheinberg, 2007; Pagan et al., 2013; Woloszyn and Sheinberg, 2009). However, we understand very little about the nature of target match signals, their behavioral relevance, and how these signals are mixed with visual representations.

The nature of the target match signal has been investigated most extensively with traditional versions of the delayed-match-to-sample (DMS) paradigm, which involves the presentation of a cue image indicating a target’s identity, followed by the presentation of a random number of distractors and then a target match (e.g. Haenny et al., 1988; Miller and Desimone, 1994; Pagan et al., 2013). During classic DMS tasks in which the cue is presented at the beginning of each trial (and the match is thus a repeat later on), IT has been reported to reflect target match information with approximately equal numbers of neurons preferring target matches versus those preferring distractors (i.e. “target match enhancement” and “target match suppression”,

respectively; Miller and Desimone, 1994; Pagan et al., 2013). Upon observing that target match suppression also follows from the repetition of distractor images within a trial, and thus cannot account for a signal that corresponds to a “target match” behavioral report, some have speculated that target match enhancement alone reflects the signal used to make behavioral judgments about whether a target match is present (Miller and Desimone, 1994). Others have proposed that the responses of both target match enhanced and suppressed subpopulations are incorporated to make behavioral judgments, particularly when a task requires disambiguating changes in firing rate due to the presence of a target match from other factors that impact overall firing rate, such as stimulus contrast (Engel and Wang, 2011). Notably, no study to date has produced compelling evidence that either IT target match enhancement or suppression accounts for (or correlates with) behavioral reports (e.g. on error trials).

Another limitation of the traditional DMS paradigm is that the cue image tends to be an exact copy of the target match, whereas real-world object search involves searching for an object that can appear at different positions, sizes and background contexts. One DMS study examined IT neural responses during this type of object variation and reported the existence of target match signals under these conditions (Leuschow et al., 1994). However, we still do not understand how IT target match signals are intermingled with IT invariant object representations of the currently-viewed scene. One intriguing proposal (Fig 1) suggests how visual and target match signals might be multiplexed to minimize the interference between them. That is, insofar as visual representations of different images are reflected as distinct patterns of spikes across the IT population (reviewed by DiCarlo et al., 2012), this translates into a population representation in which visual information is reflected by the population vector angle (Fig 1, ‘Visual modulation’). If the introduction of target match modulation also changes population vector angles in IT, this could result in perceptual confusions about the visual scene. However, if target match modulation amounts to multiplicative rescaling of population response vector lengths, this would minimize interference when superimposing visual memories and target match representations within the same network (Fig 1, ‘Target match modulation’). The degree to which the target match signal acts in this way remains unknown.

Figure 1. *Multiplexing visual and target match representations*. Shown are the hypothetical population responses to two images, each viewed (at different times) as target matches versus as distractors, plotted as the spike count response of neuron 1 versus neuron 2. In this scenario, visual information (e.g. image or object identity) is reflected by the population response pattern, or equivalently, the angle that each population response vector points. In contrast, target match information is reflected by changes in population vector length (e.g. multiplicative rescaling). Because target match information does not impact vector angle in this hypothetical scenario, superimposing target match information in this way would mitigate the impact of intermingling target match signals within underlying perceptual representations.

To investigate the nature of the IT target match signal, its behavioral relevance, and how it intermingles with IT visual representations, we recorded neural signals in IT as monkeys performed a modified delayed-match-to-sample task in which they were rewarded for indicating when a target object appeared across changes in the objects’ position, size and background context.

Results:

<u>The invariant delayed-match-to-sample task (IDMS)</u>

To investigate the target match signal, we trained two monkeys to perform an "invariant delayed-match-to-sample" (IDMS) task that required them to report when target objects appeared across variation in the objects' positions, sizes and background contexts. In this task, the target object was held fixed for short blocks of trials (~3 minutes on average) and each block began with a cue trial indicating the target for that block (Fig 2a, "Cue trial"). Subsequent test trials always began with the presentation of a distractor and on most trials this was followed by 0-5 additional distractors (for a total of 1-6 distractor images) and then an image containing the target match (Fig 2a, "Test trial"). The monkeys' task required them to fixate during the presentation of distractors and make a saccade to a response dot on the screen following target match onset to receive a reward. To minimize the predictability of the match appearing as a trial progressed, on a small subset of the trials the match did not appear and the monkey was rewarded for maintaining fixation. Our experimental design differs from other classic DMS tasks (e.g. Miller and Desimone, 1994; Pagan et al., 2013) in that it does not incorporate a cue at the beginning of each test trial, to better mimic real-world object search conditions in which target matches are not repeats of the same image presented shortly before.

Figure 2. ***The invariant delayed-match-to-sample task.*** **a)** Monkeys performed an invariant delayed-match-to-sample task. Each block (~3 minutes in duration) began with a cue trial indicating the target object for that block. On subsequent trials, monkeys initiated a trial by fixating on a small dot. After a 250 ms delay, a random number (1-7) of distractors were presented, and on most trials, this was followed by the target match. Monkeys were required to maintain fixation throughout the distractors and make a saccade to a response dot within a window 75 - 600 ms following the onset of the target match to receive a reward. In cases where the target match was presented for 400 ms and the monkey had still not broken fixation, a distractor stimulus was immediately presented. **b)** The experiment included 4 objects presented at each of 5 identity-preserving transformations ("up", "left", "right", "big", "small"), for 20 images in total. In any given block, 5 of the images were presented as target matches and 15 were distractors. **c)** The complete experimental design included looking "at" each of 4 objects, each presented at 5 identity-preserving transformations (for 20 images in total), viewed in the context of looking "for" each object as a target. In this design, target matches (highlighted in gray) fall along the diagonal of each "looking at" / "looking for" transformation slice. **d)** Percent correct for each monkey, calculated based on both misses and false alarms (but disregarding fixation breaks). Mean percent correct is plotted as a function of the position of the target match in the trial. Error bars (SEM) reflect variation across the 20 experimental sessions. **e)** Histograms of reaction times during correct trials (ms after stimulus onset) during the IDMS task for each monkey, with means indicated by arrows and labeled.

Our experiment included a fixed set of 20 images, including 4 objects presented at each of 5 transformations (Fig 2b). Our goal in selecting these specific images was to make the task of classifying object identity challenging for the IT population and these specific transformations were built on findings from our previous work (Rust and DiCarlo, 2010). In any given block (e.g. a squirrel target block), a subset of 5 of the images would be considered target matches and the

remaining 15 would be distractors (Fig 2b). Our full experimental design amounted to 20 images (4 objects presented at 5 identity-preserving transformations), all viewed in the context of each of the 4 objects as a target, resulting in 80 experimental conditions (Fig 2c). In this design, "target matches" fall along the diagonals of each looking at / looking for matrix slice (where "slice" refers to a fixed transformation; Fig 2c, gray). For each condition, we collected at least 20 repeats on correct trials. Monkeys generally performed well on this task (Fig 2d; mean percent correct monkey 1 = 96%; monkey 2 = 87%). Their mean reaction times (computed as the time their eyes left the fixation window relative to the target match stimulus onset) were 332 ms and 364 ms (Fig 2e).

As two monkeys performed this task, we recorded neural activity in IT using 24-channel probes. We performed two types of analyses on these data. The first type of analysis was performed on the data recorded simultaneously across units within a single recording session (n=20 sessions, including 10 sessions from each monkey). The second type of analysis was performed on data that was concatenated across different sessions to create a pseudopopulation after screening for units based on their stability, isolation, and task modulation (see Methods; n=204 units in total, including 108 units from monkey 1 and 96 units from monkey 2). For all but one of our analyses (Fig 3b), we counted spikes in a window that started 80 ms following stimulus onset (to allow stimulus-evoked responses time to reach IT) and ended at 250 ms, which was always before the monkeys' reaction times on these trials. For all but two of our analyses (Fig 5, 6c), the data are extracted from trials with correct responses.

<u>Target match signals were reflected in IT during the IDMS task</u>

As is typical of IT and other high-level brain areas, we encountered a heterogeneous diversity of units with regard to their tuning to different aspects of the IDMS task. Figure 3a depicts the responses of four example units, plotted as five slices through our experimental design matrix (Fig 2c), where each slice corresponds to viewing each of the four objects at a fixed transformation ('Looking AT') in the context of searching for each of the four objects as a target ('Looking FOR'). Different types of task modulation produce distinct structure in these response matrices. Visual modulation translates to vertical structure, (e.g. looking at the same image while looking for different things) whereas target modulation translates to horizontal structure (e.g. looking for the same object while looking at different things). In contrast, target match modulation is reflected as a differential response to the same images presented as target matches (which fall along the diagonal of each slice) versus distractors (which fall off the diagonal), and thus manifests as diagonal structure in each slice.

The first example unit (Fig 3a, 'visual, selective') only responded to one image (object 3 presented in the "big" transformation) and was unaffected by target identity. In contrast, the second example unit ('Fig 3a, 'visual, invariant') responded fairly exclusively to one object, but did so across four of the five transformations (all but "small"). This unit also had modest target match modulation, reflected as a larger response to its preferred object (object 1) when it was a distractor (i.e. when searching for targets 2-4) as compared to when it was a target (i.e. when searching for target 1). In other words, this unit exhibited target match suppression. The third example unit ("Fig 3a, 'one-object target match detector') consistently responded with a high firing rate to object 3 presented as a target match across all transformations, but not to other objects presented as target matches. This unit thus exhibited a form of target match enhancement that was selective for object identity. The fourth example unit ("Fig 3a, 'four-object

target match detector') responded in a compelling way with a higher firing rate response to nearly any image (any object at any transformation) presented as a target match as compared to as a distractor, or equivalently target match enhancement that was invariant to object identity. Given that the IDMS task requires an eye movement in response to images presented as target matches and fixation to the same images presented as distractors, this unit reflects something akin to the solution to the monkeys' task.

Figure 3. *Quantifying modulation in IT during the IDMS task.* **a)** The response matrices corresponding to four example IT units, plotted as the average response to five slices through the experimental design, where each slice (a 4x4 matrix) corresponds to viewing each of four objects ('Looking AT') in the context of each of four objects as a target ('Looking FOR'), at one transformation ('Big', 'Left', 'Right', 'Small', 'Up'). To compute these responses, spikes were counted in a window 80 –250 ms following stimulus onset, averaged across 20 repeated trials, and rescaled from the minimum (black) to maximum (white) response across all 80 conditions. **b)** Firing rate modulations were parsed into constituent types, where modulation was quantified in units of standard deviation around each unit's grand mean spike count (see Results). The evolution of average modulation magnitudes (across all the units for each animal; monkey 1: n = 108, monkey 2: n = 96), shown as a function of time relative to stimulus onset. The shaded area indicates the spike count window used for subsequent analyses. **c)** Average modulation magnitudes computed using the spike count window depicted in panel b.

To quantify the amounts of these different types of modulations across the IT population, we applied a procedure that quantified different types of modulation in terms of the number of standard deviations around each unit's grand mean spike count (Pagan and Rust, 2014b). Our procedure includes a bias-correction to ensure that modulations are not over-estimated by trial variability and it is similar to a multi-way ANOVA, with important extensions (see Methods). Modulation magnitudes were computed for the types described above (e.g. visual, target identity, and target match modulation) as well as "residual" modulations that are reflected as nonlinear interactions between the visual stimulus and the target identity that are not captured by target match modulation (e.g. specific distractor conditions).

Figure 3b illustrates these modulations computed in a sliding window relative to stimulus onset and averaged across all units recorded in each monkey. As expected from a visual brain area, we found that visual modulation was robust and delayed relative to stimulus onset (Fig 3b, black). Visual modulation was considerably larger in monkey 1 as compared to monkey 2. Target match modulation (Fig 3b, red) was also (as expected) delayed relative to stimulus onset and was smaller than visual modulation, but it was well above the level expected by noise (i.e. zero) and was similar in magnitude in both animals. In contrast, a robust signal reflecting information about the target identity (Fig 3b, green) appeared before stimulus onset in monkey 1 and was weaker but also present in monkey 2, consistent with a persistent working memory representation. Note that because the IDMS task was run in blocks with a fixed target, target identity information was in fact present before the onset of each stimulus. Lastly, we found that residual modulation was also present but was smaller than target match modulation in both animals (Fig 3b, cyan). When quantified in a window positioned 80 to 250 ms following stimulus onset and computed relative to the size of the target match signal (Fig 3c), visual modulation was considerably larger than target match modulation (monkey 1: 2.9x, monkey 2: 2.0x; Fig 3c, gray), whereas the other types of modulations were smaller than target match modulation

(target modulation, monkey 1: 0.9x, monkey 2: 0.5x, Fig 3c green; residual modulation, monkey 1: 0.6x, monkey 2: 0.9x Fig 3c, cyan).

In sum, the results presented thus far verify the existence of a target match signal in IT that is on average ~40% of the size of the visual modulation.

<u>IT target match information was reflected as a highly distributed, linearly separable representation</u>

The IDMS task required monkeys to determine whether each condition (an image viewed in the context of a particular target) was a target match or a distractor. This task ultimately maps all the target match conditions onto one behavioral response (a saccade) and all the distractor conditions onto another (maintain fixation), and as such, this task can be envisioned as a two-way classification across changes in other parameters, including changes in target and image identity (Fig 4a). One question of interest is the degree to which the target match versus distractor classification can be made with a linear decision boundary (or equivalently a linear decoder) applied to the IT neural data, as opposed to requiring a nonlinear decoding scheme. In a previous study, we assessed the format of IT target match information in the context of the classic DMS task design (Pagan et al., 2016; Pagan et al., 2013) and found that while a large component was linear, a considerable nonlinear (quadratic) component existed as well.

To quantify the amount and format of target match information within IT, we began by quantifying cross-validated performance for a two-way target match versus distractor classification with a weighted linear population decoder (a Fisher Linear Discriminant, FLD). Linear decoder performance began near chance and grew as a function of population size, consistent with a robust IT target match representation (Fig 4b, white). To determine the degree to which a component of IT target match information was present in a nonlinear format that could not be accessed by a linear decoder, we measured the performance of a maximum likelihood decoder designed to extract target match information regardless of its format (combined linear and nonlinear, Pagan et al., 2016; Pagan et al., 2013, see Methods). Performance of this nonlinear decoder (Fig 4b, gray) was slightly higher but not significantly better than the linear decoder for the pooled data ($p = 0.022$), and was not consistently higher in both animals (monkey 1 $p = 0.081$; monkey 2 $p = 0.647$). These results suggest that IT target match information is reflected almost exclusively in a linearly separable format during the IDMS task, at odds with our previous reports of how IT target match information is reflected during a classic DMS task (see Discussion).

Figure 4. *IT target match information is reflected via weighted linear scheme.* **a)** The target search task can be envisioned as a two-way classification of the same images presented as target matches versus as distractors. Shown are cartoon depictions where each point depicts a hypothetical response of a population of two neurons on a single trial, and clusters of points depict the dispersion of responses across repeated trials for the same condition. Included are responses to the same images presented as target matches and as distractors. Here only 6 images are depicted but 20 images were used in the actual analysis. The dotted line depicts a hypothetical linear decision boundary. **b)** Linear (FLD) and nonlinear (maximum likelihood) decoder performance as a function of population size for a pseudopopulation of 204 units, and for each monkey individually (monkey 1: n = 108 units; monkey 2: n = 96 units.) Error bars

(SEM) reflect the variability that can be attributed to the random selection of units (for populations smaller than the full dataset) and the random assignment of training and testing trials in cross-validation. **c)** Linear (FLD) decoder performance as a function of the number of top-ranked units removed. Shaded error (SEM) reflects the variability that can be attributed to the random assignment of training and testing trials in cross-validation.

Next we wanted to better understand how target match information was distributed across the IT population. We thus performed an analyses targeted at the impact of excluding the N "best" target match units for different values of N, with the rationale that if it were the case that the majority of target match information was carried by a small subpopulation of units, performance should fall quickly when those units are excluded. For this analysis, we considered the magnitude but not the sign of the target match modulation (whereas we address questions related to parsing target match modulation by sign, or equivalently target match enhancement versus suppression, below in Figure 6). To perform this analysis, we excluded the top-ranked IT units via a cross-validated procedure (i.e. based on the training data; see Methods). Consistent with a few units that carry target match signals that are considerably stronger than the rest of the population, we found that the slope of the performance drop following the exclusion of the best units was steepest for the top 8% (n=16) ranked units, and that these units accounted for ~25% of total population performance (Fig 4c). However, it was also the case that population performance continued to decline steadily as additional units were excluded, and consequently, population performance could not be attributed to a small fraction of top-ranked units alone (Fig 4c). For example, a 50% decrement in performance required removing 27% (n=55/204) of the "best-ranked" IT population, and mean +/- SEM performance remained above chance up to the elimination of 78% (n=160/204) of top-ranked units. These results are consistent with target match signals that are strongly reflected in a few units (such as Fig 3a example unit 4), and are more modestly distributed across a large fraction of the IT population (such as Fig 3a example unit 2).

Taken together, these results suggest that IT target match information is reflected by a weighted linear scheme and that target match performance depends on signals that are broadly distributed across most of the IT population.

<u>Projections along the IT linear decoding axis reflected behavioral confusions</u>

Upon establishing that the format of IT target match information during the IDMS task was linear (on correct trials), we were interested in determining the degree to which behavioral confusions were reflected in the IT neural data. To measure this, we focused on the data recorded simultaneously across multiple units within each session, where all units observed the same errors. With this data, we trained the linear decoder to perform the same target match versus distractor classification described for Fig 4 using data from correct trials, and we measured cross-validated performance on pairs of condition-matched trials: one for which the monkey answered correctly, and the other for which the monkey made an error. On correct trials, target match performance grew with population size and reached above chance levels in populations of 24 units (Fig 5, black). On error trials, mean +/- SE of decoder performance fell below chance, and these results replicated across each monkey individually (Fig 5, white). These results establish that IT reflects behaviorally-relevant target match information insofar as

projections of the IT population response along the FLD decoding axis co-vary with the monkeys' behavior.

Figure 5. *The IT FLD linear decoder axis reflects behavioral confusions.* Linear decoder performance, applied to the simultaneously recorded data for each session, after training on correct trials and cross-validating on pairs of correct and error trials matched for condition. Error bars (SEM) reflect the variability that can be attributed to the random selection of units (for populations smaller than the full dataset) and the random assignment of training and testing trials in cross-validation. Results are shown for the data pooled across all sessions (main plot, n= 20 sessions) as well as when the sessions are parsed by those collected from each animal (monkey 1, n=10 sessions; monkey 2, n=10 sessions).

Behaviorally-relevant target match signals were reflected as combinations of target match enhancement and suppression:

As described in the introduction, the IT target match signal has largely been studied via the classic DMS paradigm (which includes the presentation of the cue at the beginning of the trial) and previous results have reported approximately balanced mixtures of target match enhancement and suppression (Miller and Desimone, 1994; Pagan et al., 2013). While some have speculated that target match enhancement alone reflects the behaviorally-relevant target match signal (Miller and Desimone, 1994), others have argued that enhancement and suppression are both behaviorally-relevant (Engel and Wang, 2011). The results presented above demonstrate that during the IDMS task, the representation of target match information is largely linear, and projections along the FLD weighted linear axis reflect behavioral confusions. To what degree does IT target match information, including the reflection of behavioral confusions, follow from units that reflect target match information with target enhancement (positive weights) as compared to target suppression (negative weights)? In our study, this question is of particular interest in light of the fact that our experimental design does not include the presentation of a cue at the beginning of each trial, and thus minimizes the degree to which target match suppression follows passively from stimulus repetition.

To investigate this question, we computed a target match modulation index for each unit as the average difference between the responses to the same images presented as target matches versus as distractors, divided by the sum of those two quantities. This index takes on positive values for target match enhancement and negative values for target match suppression. In both monkeys, this index was significantly shifted toward positive values (Fig 6a; Wilcoxon sign rank test, monkey 1: mean = 0.063 $p=8.44\times10^{-6}$; monkey 2: mean = 0.071, $p = 2.11\times10^{-7}$). Notably, while these distributions were dominated by units that reflected target match enhancement, a small fraction of IT units in both monkeys reflected statistically reliable target match suppression as well (fraction of units that were significantly target match enhanced and suppressed, respectively, monkey 1: 49.1%, 17.6%; monkey 2: 41.7%, 8.3%; bootstrap significance test, $p<0.01$).

Figure 6. *Target match signals are reflected as mixtures of enhancement and suppression.* **a)** A target match modulation index, computed for each unit by calculating the mean spike count

response to target matches and to distractors, and computing the ratio of the difference and the sum of these two values. Positive values correspond to target match enhancement whereas negative values correspond to target match suppression. Dark bars in each histogram indicate the proportions for all units (Monkey 1: n = 108; monkey 2: n = 96) whereas gray bars indicate the fractions of units whose responses to target matches versus distractors were statistically distinguishable (bootstrap significance test, $p<0.01$). Arrows indicate the distribution means. **b)** Performance of the FLD classifier for the combined population (n=204 units), computed for all units (as described for Fig 4b), target match enhanced units ("E units") or target match suppressed units ("S units"). E units and S units were identified on each iteration of the resampling procedure based on the training data (as units with positive versus negative target match modulation across training trials, respectively) and cross-validated FLD performance was computed for only those units. **c)** Performance of the FLD classifier for populations of size 24 recorded in each session when trained on correct trials and tested on condition-matched pairs of correct ("Corr.") and error ("Err.") trials (as described for Fig 5), computed for all units, E units, and S units.

To determine the degree to which target match enhanced versus target match suppressed units contributed to population target match classification performance, we computed performance of the FLD linear decoder when isolated to the target match enhanced or target match suppressed subpopulations. More specifically, we focused on the combined data (to maximize the numbers of units, particularly given small fraction that were target match suppressed), and we computed performance for variants of the FLD classifier in which the sign of modulation was computed for each unit based on the training data, and cross-validated performance was determined for either the subset of target match enhanced units or the subset of target match suppressed units. When the analysis was isolated to target match enhanced units ("E units"), performance was virtually identical to the intact population (Fig 6b, mean+/- SEM performance for all units = 90.9+/-0.02% vs. E units = 90.6+/-0.02%), consistent with target match enhancement as the primary type of modulation driving population performance. When the analysis was isolated to target match suppressed units ("S units"), performance was lower than that of the intact population but still well above chance (Fig 6b, performance for S units = 64.4+/-0.03%). This suggests that the small fraction of target match suppressed units that do exist do in fact carry reliable target match signals.

Do both E units and S units reflect behavioral confusions on error trials? To address this question, we repeated the error trial analysis described above for Figure 5, but isolated to E or S units. Specifically, we repeated the analysis presented in Figure 5 where we considered the simultaneously recorded data collected across 24 units for each session, but isolated to the E or S units as described for Figure 6b (based on the training data), and we compared cross-validated performance on condition-matched correct versus error trials. As might be expected from Figure 6b, performance of the E units classified correct trials above chance and misclassified error trials below chance at rates similar to the entire population (Fig 6c, "All units" vs. "E units"). While performance of the S units on correct trials was weaker and mean +/- SEM performance was not above chance (53.0+/- 0.04%; Fig 6c "S units, Corr."), mean +/- SEM of performance of the S units on error trials did fall slightly below chance (46.6+/-0.02%; Fig 6c, "S units, Err."). These results indicate that the reflection of behavioral confusions in the IT neural data arises primarily from the activity of E units, and while not definitive, they suggest that behavioral confusions may also be weakly reflected in S units.

Together, these results suggest that in the IDMS experiment, target match signals are dominated by target match enhancement, but a smaller, target match suppressed subpopulation exists as well. Additionally, they suggest that the reflection of behavioral confusions in IT neural responses can largely be attributed to units that are target match enhanced, but behavioral confusions are also weakly reflected in units that are target match suppressed.

<u>The IT target match representation was configured to minimize interference with IT visual representations:</u>

As a final topic of interest, we wanted to understand how the representation of target match information was multiplexed with visual representations in IT and more specifically, whether IT had a means of minimizing the potentially detrimental impact of mixing these two types of signals. One possible way to achieve this is multiplicative rescaling, as described in Figure 1. To what degree is this happening in IT? To address this question, we quantified the degree of representational similarity between the IT population response vectors corresponding to the same images presented as target matches versus as distractors using a scale-invariant measure of similarity (the Pearson correlation, reviewed by Kriegeskorte and Kievit, 2013). More specifically, we measured the Pearson correlation between pairs of population response vectors via a split-halves procedure (see Methods), and we compared the representational similarity for the same images presented as target matches versus as distractors with other benchmarks in our experiment, including the representational similarity within the same experimental condition (i.e. random splits across repeated trials), and the representation similarity for images containing different transformations of the same object as well as different objects.

Shown in Figure 7a is the representational similarity matrix corresponding to all possible pairwise combinations of the 20 images used in this experiment, and averaged across the matrices computed when the pairs of response vectors under consideration were target matches and when they were distractors (see Methods). The matrix is organized such that the five transformations corresponding to each object are grouped together. Figure 7b reorganizes the data into plots of the mean and standard error of representational similarity computed for different pairwise comparisons. As expected, we found that the representational similarity was the highest for random splits of the trials corresponding to the same images, presented under the same conditions (Fig 7b, “Same image & condition”, mean = 0.87), which can be regarded as the noise ceiling in our data. In comparison, the representational similarity was significantly lower for different transformations of the same object (Fig 7b, “Different transforms.”; mean = 0.63; $p = 2.8*10^{-17}$) as well as for different objects (Fig 7b, “Different objects”; mean = 0.53; $p=8.1*10^{-28}$). Finally, the average representational similarity for the same images presented as target matches as compared to distractors was significantly lower than the noise ceiling (“Matches versus distractors”; mean = 0.74; $p=3.4*10^{-12}$) but was significantly higher than presenting the same object under a new transformation ($p=6.7*10^{-7}$) or presenting a different object ($p=1.5*10^{-16}$). These results suggest that the impact of target match signaling in IT is not perfectly multiplicative, but also has a considerably smaller impact on the population response than changing either the transformation an object is viewed in or the object in view.

Figure 7. *Target match signaling has minimal impact on the IT visual population response.* **a)** The representational similarity matrix, computed as the average Pearson correlation between the population response vectors computed for all possible pairs of images. Correlation was

computed based on a split halves procedure of the 20 trials corresponding to each experimental condition to create two population vectors reflecting the average across 10 trials each. Shown are the average correlations, computed between images with a fixed target and averaged across all possible targets, as well as averaged across 1000 random splits. The matrix is organized such that different transformations of the same object are grouped together, in the same order as depicted in Fig 2. **b)** The average representational similarity, computed across: “Same image and condition”: different random splits of the 20 trials into two sets of 10 trials each; “Different transforms.”: images containing different transformations of the same object, computed with a fixed target identity; “Different objects”: images containing different objects, computed with a fixed target identity; “Match versus distractor”: the same image viewed as a target match as compared to as a distractor, averaged across all 9 possible distractor combinations (see Methods). Error bars (SEM) reflect variability across the 20 images.

Discussion:

Successfully finding a sought target object, such as your car keys, requires your brain to compute a target match signal that reports when a target is in view. Target match signals have been reported to exist in IT, but these signals are not well understood, particularly in the context of the real-world problem of searching for an object that can appear at different identity-preserving transformations. We recorded responses in IT as two monkeys performed a delayed-match-to-sample task in which a target object could appear at different positions, sizes, and background contexts. We found that the IT population reflected a target match representation that was largely linear, and that it reflected behavioral confusions on trials in which the monkeys made errors. IT target match signals were broadly distributed across most IT units, and while they were dominated by target match enhancement, we also found evidence for reliable target match suppression. Finally, we found that IT target match modulation was configured in such a manner as to minimally impact IT visual representations. Together, these results support the existence of a robust, behaviorally-relevant target match representation in IT that is multiplexed with IT visual representations.

Our results support the existence of a robust target match representation in IT during this task that reflects confusions on trials in which the monkeys make errors (Fig 5); this result has not been reported previously. One earlier study also explored the responses of IT neurons in the context of a DMS task in which, like ours, the objects could appear at different identity-preserving transformations (Leuschow et al., 1994), but this study did not sort neural responses based on behavior. Another study examined IT neural responses as monkeys performed a visual target search task that involved free viewing as well as image manipulation during the time of the saccade (Mruczek and Sheinberg, 2007). They reported higher firing rates in IT neurons during trial sequence that normally led to a reward (an association between a target object and a saccade to a response target) versus swap trials in which this sequence was disrupted. Another study (from our lab) used a classic DMS design reported that IT population classifications on error trials fell to chance (Pagan et al., 2013), but this study did not find evidence for significant error trial misclassifications.

IT target match signals have been investigated most extensively in IT via a classic version of the delayed-match-to-sample (DMS) paradigm where each trial begins with a visual cue indicating the identity of the target object, and this cue is often the same image as the target match (Eskandar et al., 1992; Miller and Desimone, 1994; Pagan et al., 2013). In this paradigm, approximately half of all IT neurons that differentiate target matches from distractors do so with enhanced responses to matches whereas the other half are match suppressed (Miller and Desimone, 1994; Pagan et al., 2013). Because match suppressed responses also follow from the repetition of distractors within a trial, some have speculated that the match enhanced neurons alone carry behaviorally-relevant target match information (Miller and Desimone, 1994). In general agreement with those notions, the target match signal is dominated by target match enhancement in situations where the cue and target match are presented at different locations (Chelazzi et al., 1993). Conversely, others have argued that a representation comprised exclusively of match enhanced neurons would confuse the presence of a match with modulations that evoke changes in overall firing rate, such as changes in stimulus contrast (Engel and Wang, 2011). Additionally, these authors proposed that match suppressed neurons could be used in these cases to disambiguate target match versus stimulus-induced modulation. In our experiment, the IDMS task was run in blocks containing a fixed target to minimize the

impact of passive stimulus repetition of the target match. We found evidence for units that reflected target match signals with a form of repetition suppression that was reliable and reflected behavioral confusions on error trials. However, repetition suppression was also weaker than repetition enhancement. Whether the amount of repetition suppression that we observed is sufficient for the disambiguation strategy proposed by Engel and Wang (2011) is thus unclear - because our experiment did not include variation in parameters that change overall firing rate (such as contrast), we cannot directly test it with our data.

How does the target match signal arrive in IT? Computation of the target match signal requires a comparison of the content of the currently-viewed scene with a remembered representation of the sought target. The existence of target match signals in IT could reflect the implementation of the comparison in IT itself or, alternatively, this comparison might be implemented in a higher-order brain area (such as prefrontal cortex) and fed-back to IT. While examination of the timing of the arrival of this signal in IT (which peaks at 150 ms; Fig 3b) relative to the monkeys' median reaction times (~340 ms; Fig 2e) do not rule out the former scenario, we cannot definitively distinguish between these alternative scenarios with our data.

In a previous series of reports, we investigated target match signals in the context of the classic DMS design in which target matches were repeats of cues presented earlier in the trial and each object was presented on a gray background (Pagan and Rust, 2014a; Pagan et al., 2016; Pagan et al., 2013). One of our main findings from that work was that the IT target match representation was reflected in a partially nonlinearly separable format, whereas an IT downstream projection area, perirhinal cortex, contained the same amount of target match information but in a format that was largely linearly separable. In the data we present here, we did not find evidence for a nonlinear component of the IT target match representation, reflected by higher performance of a maximum likelihood as compared to linear decoder (Fig 4b). The source of these differences is unclear. They could arise from the fact that the IDMS task requires an "invariant" visual representation of object identity, which first emerges in a linearly separable format in the brain area that we are recording from (Rust and DiCarlo, 2010), whereas in more classic forms of the DMS task, the integration of visual and target information could happen in a different manner and/or a different brain area. Alternatively, these differences could arise from the fact that during IDMS, images are not repeated within a trial, and the stronger nonlinear component revealed in DMS may be produced by stimulus repetition. Our current data cannot distinguish between these alternatives.

Our results also add to the growing literature that suggests the brain "mixes" the modulations for different task-relevant parameters within individual neurons, even at the highest stages of processing (Freedman and Assad, 2009; Kobak et al., 2016; Mante et al., 2013; Meister et al., 2013; Raposo et al., 2014; Rigotti et al., 2013; Rishel et al., 2013; Zoccolan et al., 2007). A number of explanations have been proposed to account for mixed selectivity. Some studies have documented situations in which signal mixing is an inevitable consequence of the computations required for certain tasks, such as identifying objects invariant to the view in which they appear (Zoccolan et al., 2007). Others have suggested that mixed selectivity may be an essential component of the substrate required to maintain a representation that can rapidly and flexibly switch with changing task demands (Raposo et al., 2014; Rigotti et al., 2013). Still others have maintained that broad tuning across different types of parameters is important for learning new associations (Barak et al., 2013). Our results suggest that IT mixes visual and target match information within individual units. This could reflect the fact that the comparison of visual and target match information happens within IT itself, and multiplexing is simply a byproduct of that

computation. Alternatively, if the comparison is performed elsewhere, this would reflect its feedback to IT for some unknown purpose. In either case, our results suggest that the multiplexing happens in a manner that is largely but imperfectly multiplicative (Fig 7) and thus configured to minimize interference of the visual representation when also signaling target match information.

Acknowledgements:

We thank Margot P. Wohl for her contributions to early phases of this work. This work was supported by the National Eye Institute of the National Institutes of Health (award R01EY020851), the Simons Foundation (through an award from the Simons Collaboration on the Global Brain), and the McKnight Endowment for Neuroscience.

METHODS

Experiments were performed on two adult male rhesus macaque monkeys (*Macaca mulatta)* with implanted head posts and recording chambers. All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee.

The invariant delayed-match-to-sample (IDMS) task:

All behavioral training and testing was performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, infrared video eye tracking. Stimuli were presented on an LCD monitor with an 85 Hz refresh rate using customized software (http://mworks-project.org).

As an overview, the monkeys' task required an eye movement response to a specific location when a target object appeared within a sequence of distractor images (Fig 2a). Objects were presented across variation in the objects' position, size and background context (Fig 2b). Monkeys viewed a fixed set of 20 images across switches in the identity of 4 target objects, each presented at 5 identity-preserving transformations (Fig 2c). We ran the task in short blocks (~3 min) with a fixed target before another target was pseudorandomly selected. Our design included two types of trials: cue trials and test trials (Fig 2a). Only test trials were analyzed for this report.

Trials were initiated by the monkey fixating on a red dot (0.15°) in the center of a gray screen, within a square window of ±1.5°, followed by a 250 ms delay before a stimulus appeared. Cue trials, which indicated the current target object, were presented at the beginning of each block and after three subsequent trials with incorrect responses. To minimize confusion, cue trials were designed to be distinct from test trials and began with the presentation of an image of each object that was distinct from the images used on test trials (a large version of the object presented at the center of gaze on a gray background; Fig 2a). Test trials, which are the focus of this report, always began with a distractor image, and neural responses to this image were discarded to minimize non-stationarities such as stimulus onset effects. Distractors were drawn randomly from a pool of 15 possible images within each block without replacement until each distractor was presented once on a correct trial, and the images were then re-randomized. On most trials, a random number of 1-6 distractors were presented, followed by a target match (Fig 2a). On a small fraction of trials, 7 distractors were shown, and the monkey was rewarded for fixating through all distractors. Each stimulus was presented for 400 ms (or until the monkeys' eyes left the fixation window) and was immediately followed by the presentation of the next stimulus. Following the onset of a target match image, monkeys were rewarded for making a saccade to a response target within a window of 75 – 600 ms to receive a juice reward. In monkey 1 this target was positioned 10 degrees above fixation; in monkey 2 it was 10 degrees below fixation. If 400 ms following target onset had elapsed and the monkey had not moved its eyes, a distractor stimulus was immediately presented. If the monkey continued fixating beyond the required reaction time, the trial was considered a "miss". False alarms were differentiated from fixation breaks via a comparison of the monkeys' eye movements with the characteristic pattern of eye movements on correct trials: false alarms were characterized by the eyes leaving the fixation window via its top (monkey 1) or bottom (monkey 2) outside the allowable correct

response period and traveling more than 0.5 degrees whereas fixation breaks were characterized by the eyes leaving the fixation window in any other way. Within each block, 4 repeated presentations of the 20 images were collected, and a new target object was then pseudorandomly selected. Following the presentation of all 4 objects as targets, the targets were re-randomized. At least 20 repeats of each condition were collected. Overall, monkeys performed this task with high accuracy. Disregarding fixation breaks (monkey 1: 11% of trials, monkey 2: 8% of trials), percent correct on the remaining trials was as follows: monkey 1: 96% correct, 1% false alarms, and 3% misses; monkey 2: 87% correct, 3% false alarms, and 10% misses.

Neural recording:

The activity of neurons in IT was recorded via a single recording chamber in each monkey. Chamber placement was guided by anatomical magnetic resonance images in both monkeys, and in one monkey, Brainsight neuronavigation (https://www.rogue-research.com/). The region of IT recorded was located on the ventral surface of the brain, over an area that spanned 4 mm lateral to the anterior middle temporal sulcus and 15-19 mm anterior to the ear canals. Neural activity was largely recorded with 24-channel U probes (Plexon, Inc) with linearly arranged recording sites spaced with 100 μm intervals, with a handful of units recorded with single electrodes (Alpha Omega, glass-coated tungsten). Continuous, wideband neural signals were amplified, digitized at 40 kHz and stored using the OmniPlex Data Acquisition System (Plexon). Spike sorting was done manually offline (Plexon Offline Sorter). At least one candidate unit was identified on each recording channel, and 2-3 units were occasionally identified on the same channel. Spike sorting was performed blind to any experimental conditions to avoid bias. A multi-channel recording session was included in the analysis if the animal performed the task until the completion of 20 correct trials per stimulus condition, there was no external noise source confounding the detection of spike waveforms, and the session included a threshold number of task modulated units (>4 on 24 channels). The sample size (number of units recorded) was chosen to approximately match our previous work (Pagan and Rust, 2014a; Pagan et al., 2016; Pagan et al., 2013).

For all the analyses presented in this paper except Fig 3b, we measured neural responses by counting spikes in a window that began 80 ms after stimulus onset and ended at 250 ms. On 1.9% of all correct target match presentations, the monkeys had reaction times faster than 250 ms, and those instances were excluded from analysis such that spikes were only counted during periods of fixation. When combining the units recorded across sessions into a larger pseudopopulation, we screened for units that met three criteria. First, units had to be modulated by our task, as quantified by a one-way ANOVA applied to our neural responses (80 conditions * 20 repeats) with $p < 0.01$. Second, we applied a loose criterion on recording stability, as quantified by calculating the variance-to-mean for each unit (computed by fitting the relationship between the mean and variance of spike count across the 80 conditions), and eliminating units with a variance-to-mean ratio > 5. Finally, we applied a loose criterion on unit recording isolation, quantified by calculating the signal-to-noise ratio (SNR) of the waveform (as the difference between the maximum and minimum points of the average waveform, divided by twice the standard deviation across the differences between each waveform and the mean waveform), and excluding (multi)units with an SNR < 2. This yielded a pseudopopulation of 204 units (of 563 possible units), including 108 units from monkey 1 and 96 units from monkey 2.

Quantifying single-unit modulation magnitudes:

To quantify the degree to which individual units were modulated by different types of task parameters (Fig 3b-d), we applied a bias-corrected procedure described in detail by (Pagan and Rust, 2014b) and summarized here. Our measure of modulation is similar to a multi-way ANOVA, with important extensions. Specifically, a two-way ANOVA applied to a unit's responses (configured into a matrix of 4 targets * 20 images * 20 trials for each condition) would parse the total response variance into two linear terms, a nonlinear interaction term, and an error term. We make 3 extensions to the ANOVA analysis. First, an ANOVA returns measures of variance (in units of spike counts squared) whereas we compute measures of standard deviation (in units of spike count) such that our measures of modulation are intuitive (e.g., doubling firing rates causes signals to double as opposed to quadruple). Second, while the linear terms of the ANOVA map onto our "visual" and "target identity" modulations (after squaring), we split the ANOVA nonlinear interaction term into two terms, including target match modulation (i.e. Fig 2c gray versus white) and all other nonlinear "residual" modulation. This parsing is essential, as target match modulation corresponds to the signal for the IDMS task whereas the other types of modulations are not. Finally, raw ANOVA values are biased by trial-by-trial variability (which the ANOVA addresses by computing the probability that each term is higher than chance given this noise) whereas our measures of modulation are bias-corrected to provide an unbiased estimate of modulation magnitude.

The procedure begins by developing an orthonormal basis of 80 vectors designed to capture all types of modulation with intuitive groupings. The number of each type is imposed by the experimental design. This basis $\boldsymbol{b}$ included vectors $\boldsymbol{b}_i$ that reflected 1) the grand mean spike count across all conditions ($\boldsymbol{b}_1$, 1 dimension), 2) whether the object in view was a target or a distractor ($\boldsymbol{b}_2$, 1 dimension), 3) visual image identity ($\boldsymbol{b}_3 - \boldsymbol{b}_{21}$, 19 dimensions), 4) target object identity ($\boldsymbol{b}_{22} - \boldsymbol{b}_{24}$, 3 dimensions), and 5) "residual", nonlinear interactions between target and object identity not captured by target match modulation ($\boldsymbol{b}_{25} - \boldsymbol{b}_{80}$, 56 dimensions). A Gram-Schmidt process was used to convert an initially designed set of vectors into an orthonormal basis.

Because this basis spans the space of all possible responses for our task, each trial-averaged vector of spike count responses to the 80 experimental conditions $\boldsymbol{R}$ can be re-expressed as a weighted sum of these basis vectors. To quantify the amounts of each type of modulation reflected by each unit, we began by computing the squared projection of each basis vector $\boldsymbol{b}_i$ and $\boldsymbol{R}$. An analytical bias correction, described and verified in (Pagan and Rust, 2014b), was then subtracted from this value:

(8) $$w_i^2 = (\boldsymbol{R} \cdot \boldsymbol{b}_i^T)^2 - \frac{\sigma_t^2 \cdot (\boldsymbol{b}_i^T)^2}{m}$$

where σ_t^2 indicates the trial variance, averaged across conditions (n=80), and where m indicates the number of trials (m=20). When more than one dimension existed for a type of modulation, we summed values of the same type. Next, we applied a normalization factor (1/(n-1)) to convert these summed values into variances. Finally, we computed the square root of these quantities

to convert them into modulation measures that reflected the number of spike count standard deviations around each unit's grand mean spike count.

Target match modulation was thus computed as:

(9) $\sigma_{TM} = \sqrt{\frac{1}{n-1} \cdot w_2^2}$

visual modulation was computed as:

(10) $\sigma_{Vis} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=3}^{21} w_i^2}$

target identity modulation was computed as:

(11) $\sigma_{TI} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=22}^{24} w_i^2}$

and residual modulation was computed as:

(12) $\sigma_{res} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=25}^{80} w_i^2}$

When estimating modulation population means (Fig 3b, c), the bias-corrected squared values were averaged across units before taking the square root. Because these measures were not normally distributed, standard error about the mean was computed via a bootstrap procedure. On each iteration of the bootstrap (across 1000 iterations), we randomly sampled values from the modulation values for each unit in the population, with replacement. Standard error was computed as the standard deviation across the means of these newly created populations.

To quantify the sign of the modulation corresponding to whether an image was presented as a target match versus as a distractor (Fig 6), we calculated a target match modulation index for each unit by computing its mean spike count response to target matches and to distractors, and computing the ratio of their difference and their sum.

Population performance:

To determine the performance of the IT population at classifying target matches versus distractors, we applied two types of decoders: a Fisher Linear Discriminant (a linear decoder) and Maximum Likelihood decoder (a nonlinear decoder) using approaches that are described previously in detail (Pagan et al., 2013) and are summarized here.

When applied to the pseudopopulation data (Fig 4b, Fig 6b), all decoders were cross-validated with the same resampling procedure. On each iteration of the resampling, we randomly shuffled the trials for each condition and for each unit, and (for numbers of units less than the full population size) randomly selected units. On each iteration, 18 trials from each condition were

used for training the decoder, 1 trial was used to determine a value for regularization, and 1 trial from each condition was used for cross-validated measurement of performance.

To ensure that decoder performance was not biased by unequal numbers of target matches and distractors, on each iteration of the resampling we included 20 target match conditions and 20 (of 60 possible) distractor conditions. Each set of 20 distractors was selected to span all possible combinations of mismatched object and target identities (e.g. objects 1, 2, 3, 4 paired with targets 4, 3, 2, 1), of which there are 9 possible sets. To compute proportion correct, a mean performance value was computed on each resampling iteration by averaging binary performance outcomes across the 9 possible sets of target matches and distractors, each which contained 40 test trials. Mean and standard error of performance was computed as the mean and standard deviation of performance across 1000 resampling iterations. Standard error thus reflected the variability due to the specific trials assigned to training and testing and, for populations smaller than the full size, the specific units chosen.

Fisher Linear Discriminant:

The general form of a linear decoding axis is:

(1) $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b,$

where **w** is an N-dimensional vector (where N is the number of units) containing the linear weights applied to each unit, and b is a scalar value. We fit these parameters using a Fisher Linear Discriminant (FLD), where the vector of linear weights was calculated as:

(2) $\boldsymbol{w} = \Sigma^{-\mathbf{1}}(\mu_1 - \mu_2)$

and b was calculated as:

(3) $b = \boldsymbol{w} \cdot \frac{1}{2}(\mu_1 + \mu_2) = \frac{1}{2}{\mu_1}^T\Sigma^{-\mathbf{1}}\mu_1 - \frac{1}{2}{\mu_2}^T\Sigma^{-\mathbf{1}}\mu_2$

Here μ_1 *and* μ_2 are the means of the two classes (target matches and distractors, respectively) and the mean covariance matrix is calculated as:

(4) $\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$

where Σ_1 and Σ_2 are the regularized covariance matrices of the two classes. These covariance matrices were computed using a regularized estimate equal to a linear combination of the sample covariance and the identity matrix I (Pagan et al., 2016):

(5) $\Sigma_i = \gamma\, \Sigma_i + (1 - \gamma) \cdot I$

We determined γ by exploring a range of values from 0.01 to 0.99, and we selected the value that maximized average performance across all iterations, measured with the cross-validation "regularization" trials set aside for this purpose (see above). We then computed performance for that value of γ with separately measured "test" trials, to ensure a fully cross-validated measure. Because this calculation of the FLD parameters incorporates the off-diagonal terms of the

covariance matrix, FLD weights are optimized for both the information conveyed by individual units as well as their pairwise interactions.

To compare FLD performance on correct versus error trials (Fig 5, 6c), we used the same methods described above with the following modifications. First, the analysis was applied to the simultaneously recorded data within each session, and the correlation structure on each trial was kept intact on each resampling iteration. Second, when more than 24 units were available, a subset of 24 units were selected as those with the most task modulation, quantified via the p-value of a one-way ANOVA applied to each unit's responses (80 conditions * 20 repeats). Finally, on each resampling iteration, each error trial was randomly paired with a correct trial of the same condition and cross-validated performance was performed exclusively for these pairs of correct and error responses. As was the case for the pseudopopulation analysis, training was performed exclusively on correct trials. A mean performance value was computed on each resampling iteration by averaging binary performance outcomes across all possible error trials and their condition-matched correct trial pairs, and averaging across different recording sessions. Mean and standard error of performance was computed as the mean and standard deviation of performance across 100 resampling iterations. Standard error thus reflected error in a manner similar to the pseudopopulation analysis - the variability due to the specific trials assigned to training and testing and, for populations smaller than the full size, the specific units chosen.

In the case of the ranked-FLD (Fig 4c), all units were considered on each resampling iteration, and weights were computed for each unit (with the training data) as described by Equation 2. Weights were then ranked by their magnitude (the absolute values of the signed quantities) and the top N units were selected for different population size N. Finally, both the weights and the threshold were recalculated before cross-validated testing with the training data. In the case of the signed versions of the FLD (which isolated the analysis to target matched enhanced or suppressed units, Fig 6b-c), the process was similar in that all units were considered on each resampling iteration and weights were computed for each unit (with the training data) as described by Equation 2. Weights were then isolated to all of those that were positive "E units" or all that were negative "S units". Finally, the weights and the threshold were recalculated before cross-validated testing with the training data.

Maximum likelihood decoder:

As a measure of total IT target match information (combined linear and nonlinear), we implemented the maximum likelihood decoder (Fig 4b) introduced in our previous work (Pagan et al., 2016; Pagan et al., 2013). We began by using the set of training trials to compute the average response r_{uc} of each unit u to each of the 40 conditions c. We then computed the likelihood that a test response k was generated from a particular condition as a Poisson-distributed variable:

(7) $$lik_{u,c}(k) = \frac{(r_{uc})^{k} \cdot e^{-r_{uc}}}{k!}$$

The likelihood that a population response vector was generated in response to each condition was then computed as the product of the likelihoods of the individual units. Next, we computed the likelihood that each test vector arose from the category target match as compared to the

category distractor as the product of the likelihoods across the conditions within each category. We assigned the population response to the category with the maximum likelihood, and we computed performance as the fraction of trials in which the classification was correct based on the true labels of the test data.

Representational similarity:

To compute measures of the representational similarity between pairs of population response vectors, the 20 repeated trials for each (of 80) experimental conditions were randomly split into two sets of 10 trials, and the average population response vector was computed. To obtain measures of the noise ceiling, Pearson correlation was computed between many random splits of the data for each of the 80 conditions. The mean across 1000 random splits was computed for each condition and the values were averaged across the splits as well as the 4 target conditions for each image, resulting in 20 correlations values (1 for each image). Fig 7b “Same image and condition” depicts the mean and standard deviation across the 20 images. Measures of the representational similarity between different conditions were computed in a comparable way, by also selecting 10 (of 20) trials before computing the mean population response vectors. To measure the representational similarity between the same objects presented at different transformations, Pearson correlation was computed for all possible pairs of the 5 images corresponding to each object under the conditions of a fixed target. A mean value was computed as the average across 1000 random splits and the pairwise comparison between 1 image and other images containing the same object for each of 20 images, and Fig 7b “Different transforms.” depicts the mean and standard deviation across the 20 images. Fig 7b “Different objects” was computed in a similar manner, but for all possible pairs of one image and the images containing other objects. Finally, Fig 7b “Match versus distractor” was computed in a similar manner, but for all possible pairs of one image presented as a target match (viewing an image while looking for that object as a target) and the three distractor conditions (the three other targets).

Statistical tests:

When comparing performance between the FLD and maximums likelihood classifier (Fig 4b), we reported *P* values as an evaluation of the probability that differences were due to chance. We calculated *P* values as the fraction of resampling iterations on which the difference was flipped in sign relative to the actual difference between the means of the full data set (for example, if the mean of decoding measure 1 was larger than the mean of decoding measure 2, the fraction of iterations in which the mean of measure 2 was larger than the mean of measure 1).

When evaluating whether each unit had a statistically different response to target matches as compared to distractors (Fig 6a, light bars), we recomputed each unit’s modulation index by resampling trials with replacement on n = 1000 resampling iterations. A unit was considered as statistically significant if its resampled modulation indices were flipped in sign from the unit’s actual modulation index less than 0.01% of the resampling iterations. When evaluating whether

the single unit modulation indices (Fig 6a) were significantly different from zero, we reported P values as computed by a Wilcoxon sign rank test.

When comparing the representational similarity of different groupings of the IT population response (Fig 7b), we computed a mean Pearson correlation value for each of the 20 images (as described above), and reported P values as the probability that the observed differences in means across the 20 images were due to chance via a two-sample t-test.

References:

Barak, O., Rigotti, M., and Fusi, S. (2013). The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. J Neurosci *33*, 3844-3856.

Bichot, N.P., Rossi, A.F., and Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. Science *308*, 529-534.

Chelazzi, L., Duncan, J., Miller, E.K., and Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. J Neurophysiology *80*, 2918-2940.

Chelazzi, L., Miller, E.K., Duncan, J., and Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. Nature *363*, 345-347.

Chelazzi, L., Miller, E.K., Duncan, J., and Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. Cereb Cortex *11*, 761-772.

DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415-434.

Engel, T.A., and Wang, X.J. (2011). Same or different? A neural circuit mechanism of similarity-based pattern match decision making. J Neurosci *31*, 6982-6996.

Eskandar, E.N., Richmond, B.J., and Optican, L.M. (1992). Role of inferior temporal neurons in visual memory I. Temporal encoding of information about visual images, recalled images, and behavioral context. Journal of Neurophysiology *68*, 1277-1295.

Freedman, D.J., and Assad, J.A. (2009). Distinct encoding of spatial and nonspatial visual information in parietal cortex. J Neurosci *29*, 5671-5680.

Gibson, J.R., and Maunsell, J.H.R. (1997). The sensory modality specificity of neural activity related to memory in visual cortex. Journal of Neurophysiology *78*, 1263-1275.

Haenny, P.E., Maunsell, J.H., and Schiller, P.H. (1988). State dependent activity in monkey visual cortex. II. Retinal and extraretinal factors in V4. Exp Brain Res *69*, 245-259.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C.E., Kepecs, A., Mainen, Z.F., Qi, X.L., Romo, R., Uchida, N., and Machens, C.K. (2016). Demixed principal component analysis of neural population data. Elife *5*.

Kosai, Y., El-Shamayleh, Y., Fyall, A.M., and Pasupathy, A. (2014). The role of visual area V4 in the discrimination of partially occluded shapes. J Neurosci *34*, 8570-8584.

Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci *17*, 401-412.

Leuschow, A., Miller, E.K., and Desimone, R. (1994). Inferior temporal mechanisms for invariant object recognition. Cerebral Cortex *5*, 523-531.

Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78-84.

Maunsell, J.H., Sclar, G., Nealey, T.A., and DePriest, D.D. (1991). Extraretinal representations in area V4 in the macaque monkey. Vis Neurosci *7*, 561-573.

Meister, M.L., Hennig, J.A., and Huk, A.C. (2013). Signal multiplexing and single-neuron computations in lateral intraparietal area during decision-making. J Neurosci *33*, 2254-2267.

Miller, E.K., and Desimone, R. (1994). Parallel neuronal mechanisms for short-term memory. Science *263*, 520-522.

Mruczek, R.E., and Sheinberg, D.L. (2007). Activity of inferior temporal cortical neurons predicts recognition choice behavior and recognition time during visual search. J Neurosci *27*, 2825-2836.

Pagan, M., and Rust, N.C. (2014a). Dynamic target match signals in perirhinal cortex can be explained by instantaneous computations that act on dynamic input from inferotemporal cortex. J Neurosci *34*, 11067-11084.

Pagan, M., and Rust, N.C. (2014b). Quantifying the signals contained in heterogeneous neural responses and determining their relationships with task performance. J Neurophysiol *112*, 1584-1598.

Pagan, M., Simoncelli, E.P., and Rust, N.C. (2016). Neural Quadratic Discriminant Analysis: Nonlinear Decoding with V1-Like Computation. Neural Comput, 1-29.

Pagan, M., Urban, L.S., Wohl, M.P., and Rust, N.C. (2013). Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. Nature neuroscience *16*, 1132-1139.

Raposo, D., Kaufman, M.T., and Churchland, A.K. (2014). A category-free neural population supports evolving demands during decision-making. Nature neuroscience *17*, 1784-1792.

Rigotti, M., Barak, O., Warden, M.R., Wang, X.J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585-590.

Rishel, C.A., Huang, G., and Freedman, D.J. (2013). Independent category and spatial encoding in parietal cortex. Neuron *77*, 969-979.

Rust, N.C., and DiCarlo, J.J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. J Neurosci *30*, 12978-12995.

Woloszyn, L., and Sheinberg, D.L. (2009). Neural dynamics in inferior temporal cortex during a visual working memory task. J Neurosci *29*, 5494-5507.

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J.J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J Neurosci *27*, 12292-12307.

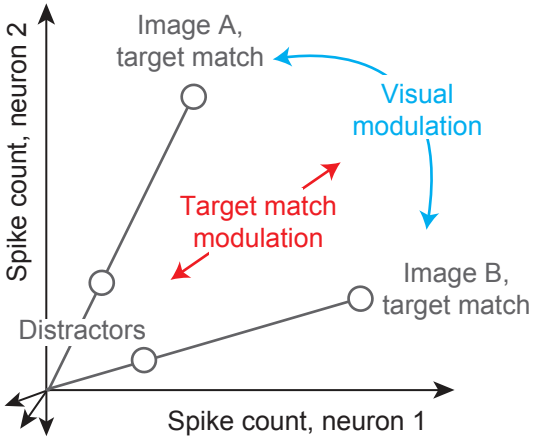
Spike count, neuron 2
Image A,
target match
Visual
modulation
Target match
modulation
Image B,
target match
Distractors
Spike count, neuron 1

Figure 1

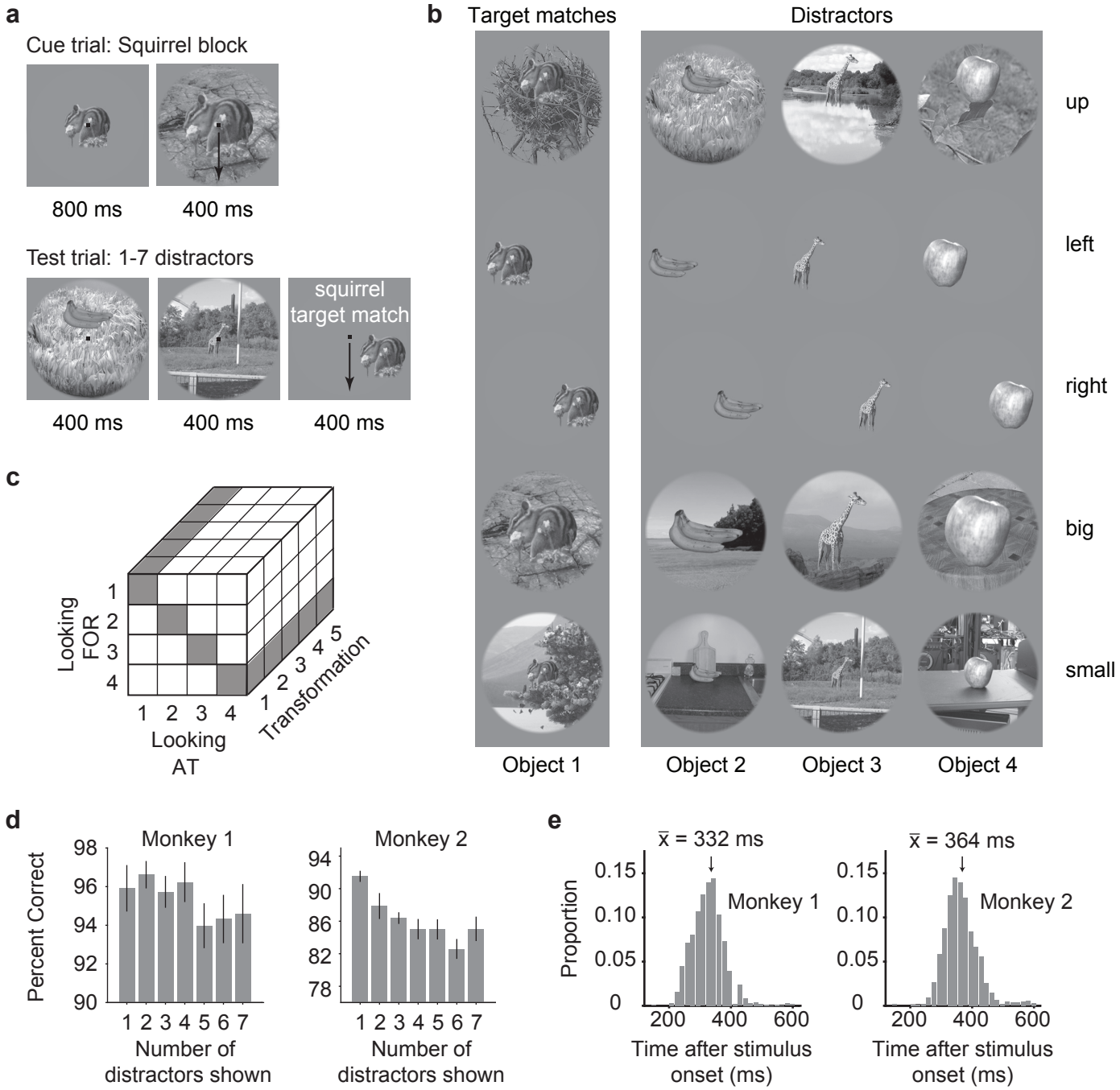
a
Cue trial: Squirrel block
800 ms
400 ms
Test trial: 1-7 distractors
squirrel target match
400 ms
400 ms
400 ms
b
Target matches
Distractors
up
left
right
big
small
Object 1
Object 2
Object 3
Object 4
c
Looking FOR
Looking AT
Transformation
d
Monkey 1
Monkey 2
Percent Correct
Number of distractors shown
e
x̄ = 332 ms
x̄ = 364 ms
Monkey 1
Monkey 2
Proportion
Time after stimulus onset (ms)

Figure 2

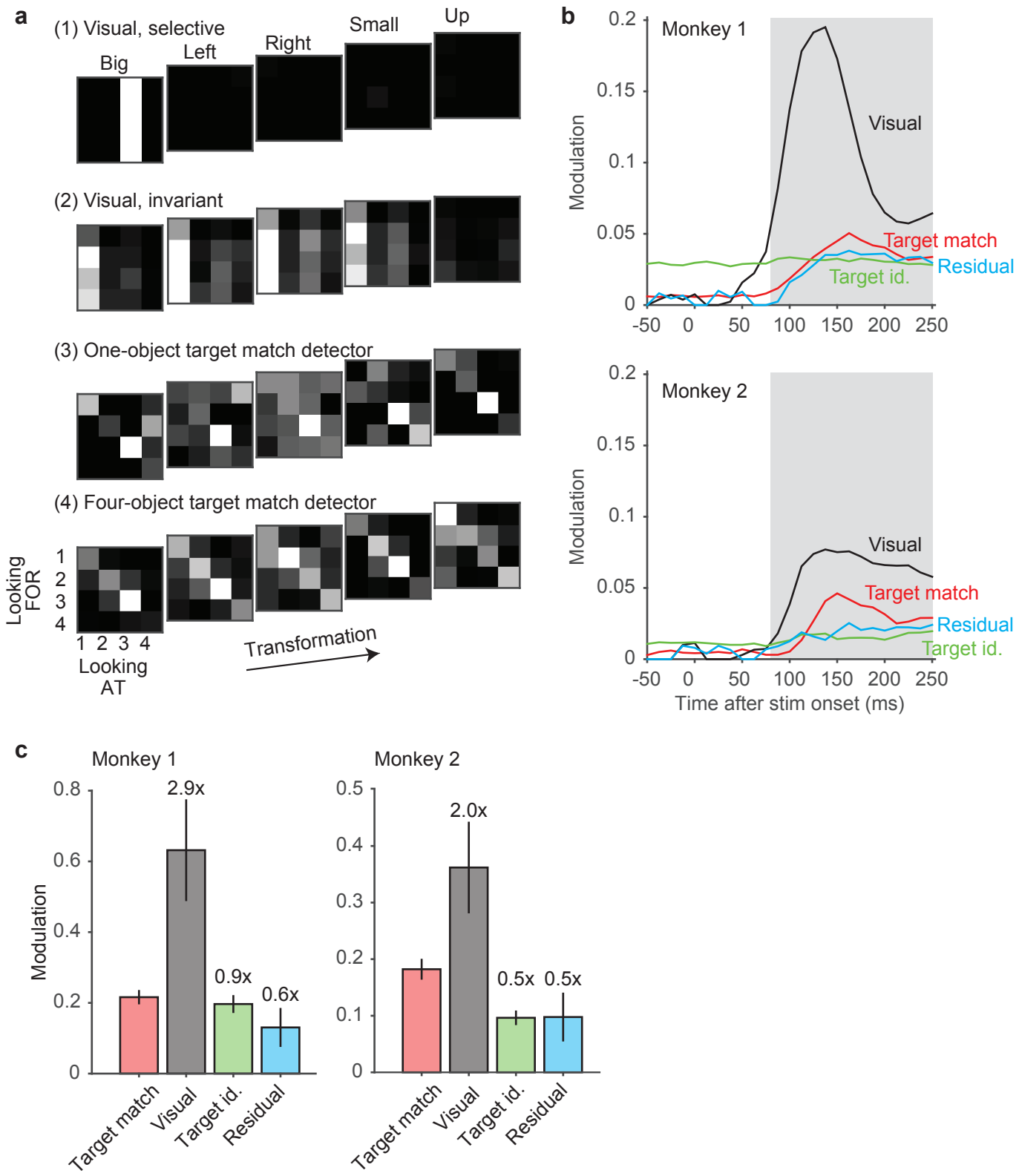
a
(1) Visual, selective
Big
Left
Right
Small
Up
(2) Visual, invariant
(3) One-object target match detector
(4) Four-object target match detector
Looking FOR
1
2
3
4
1 2 3 4
Looking AT
Transformation
b
Monkey 1
Modulation
Visual
Target match
Residual
Target id.
Monkey 2
Time after stim onset (ms)
c
Monkey 1
Monkey 2
2.9x
0.9x
0.6x
2.0x
0.5x
0.5x
Target match
Visual
Target id.
Residual

Figure 3

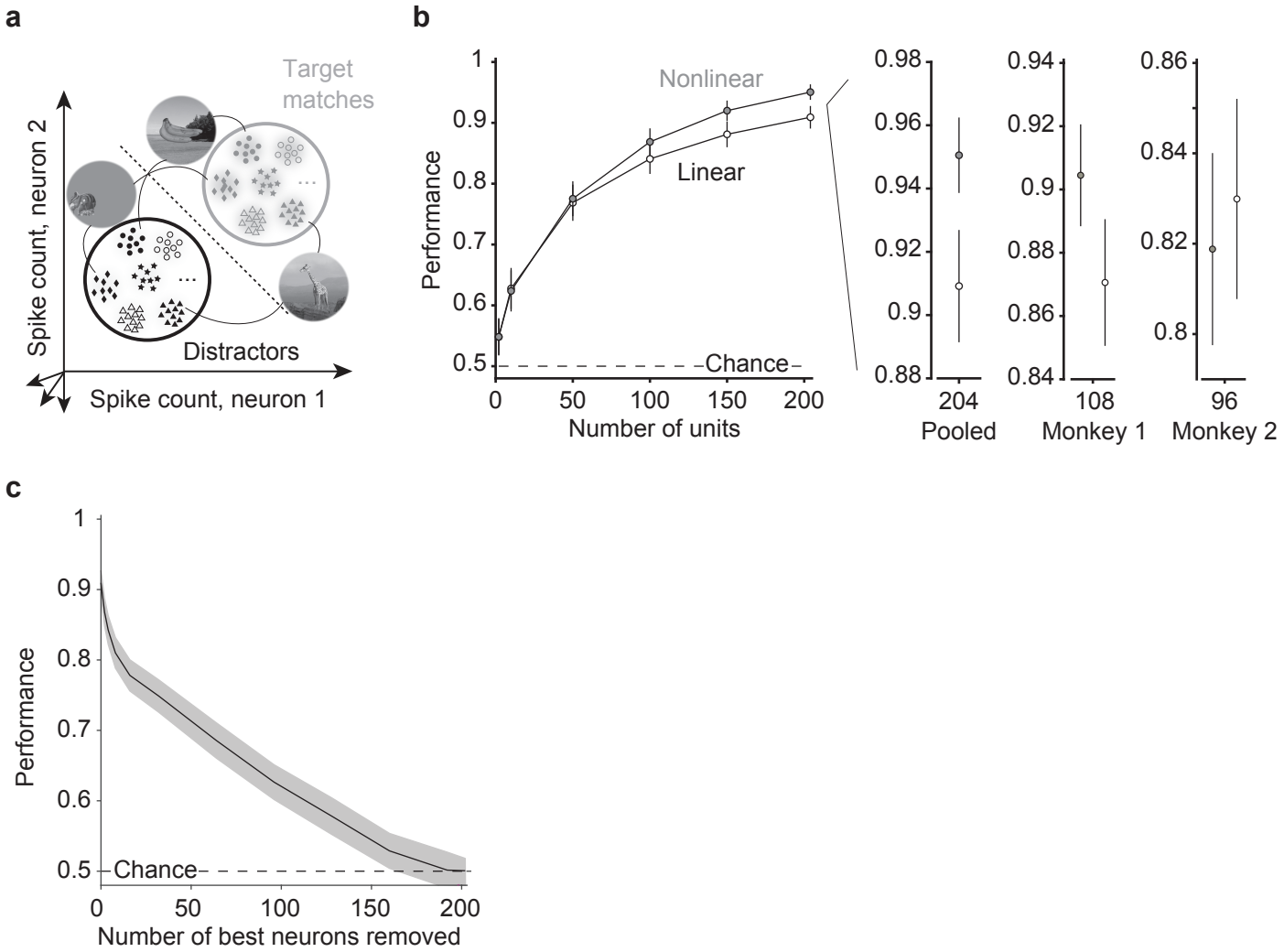
a
Target
matches
Spike count, neuron 2
Distractors
Spike count, neuron 1
b
Nonlinear
Linear
Performance
Chance
Number of units
Pooled
Monkey 1
Monkey 2
c
Performance
Chance
Number of best neurons removed

Figure 4

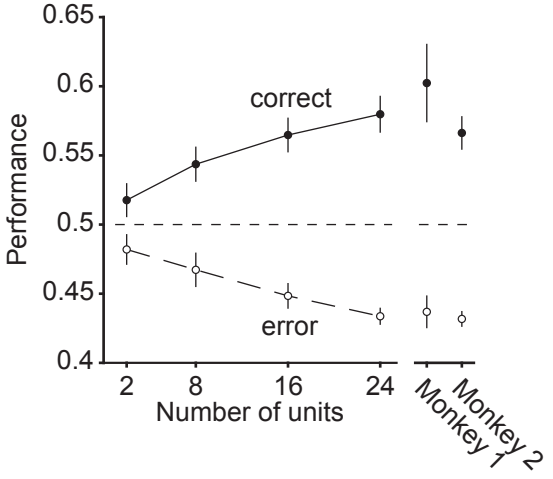
0.65
0.6
0.55
0.5
0.45
0.4
Performance
correct
error
2
8
16
24
Number of units
Monkey 1
Monkey 2

Figure 5

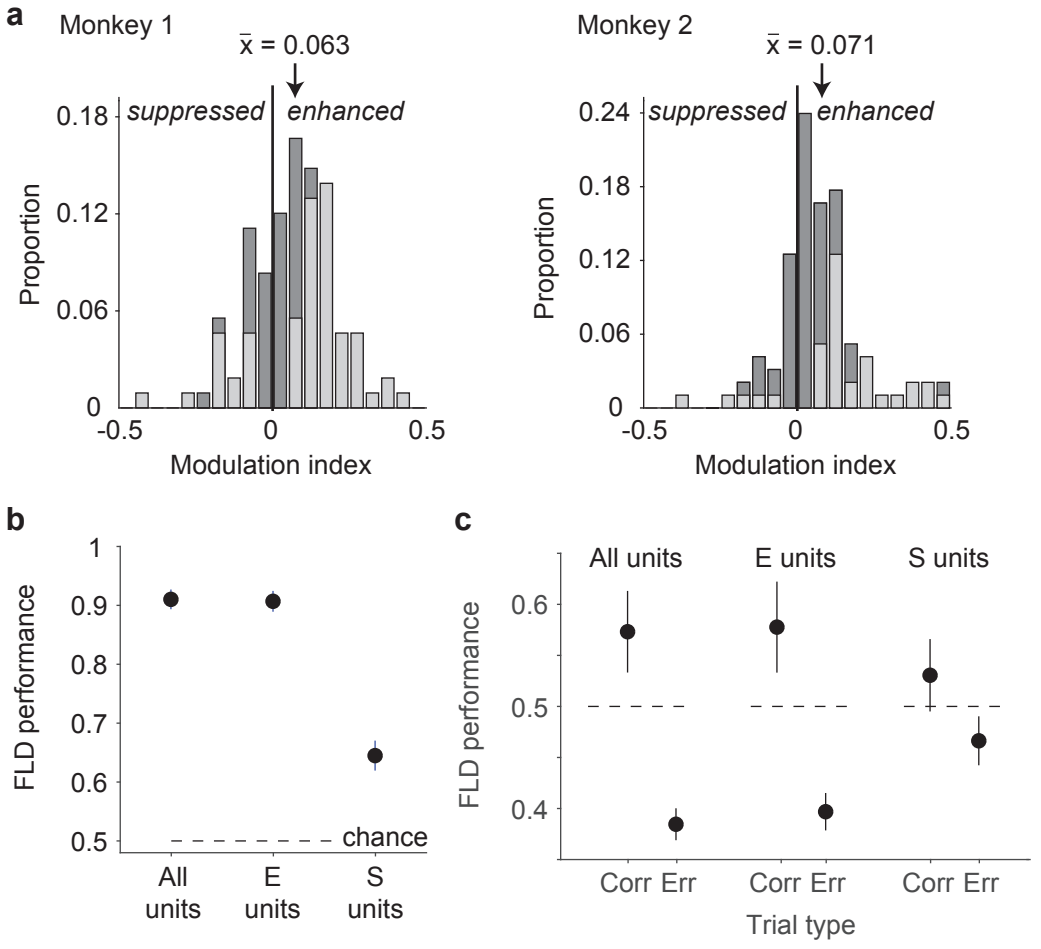
a
Monkey 1
x̄ = 0.063
suppressed
enhanced
Proportion
0.18
0.12
0.06
0
-0.5
0
0.5
Modulation index
Monkey 2
x̄ = 0.071
suppressed
enhanced
Proportion
0.24
0.18
0.12
0.06
0
-0.5
0
0.5
Modulation index
b
FLD performance
1
0.9
0.8
0.7
0.6
0.5
chance
All units
E units
S units
c
All units
E units
S units
FLD performance
0.6
0.5
0.4
Corr Err
Corr Err
Corr Err
Trial type

Figure 6

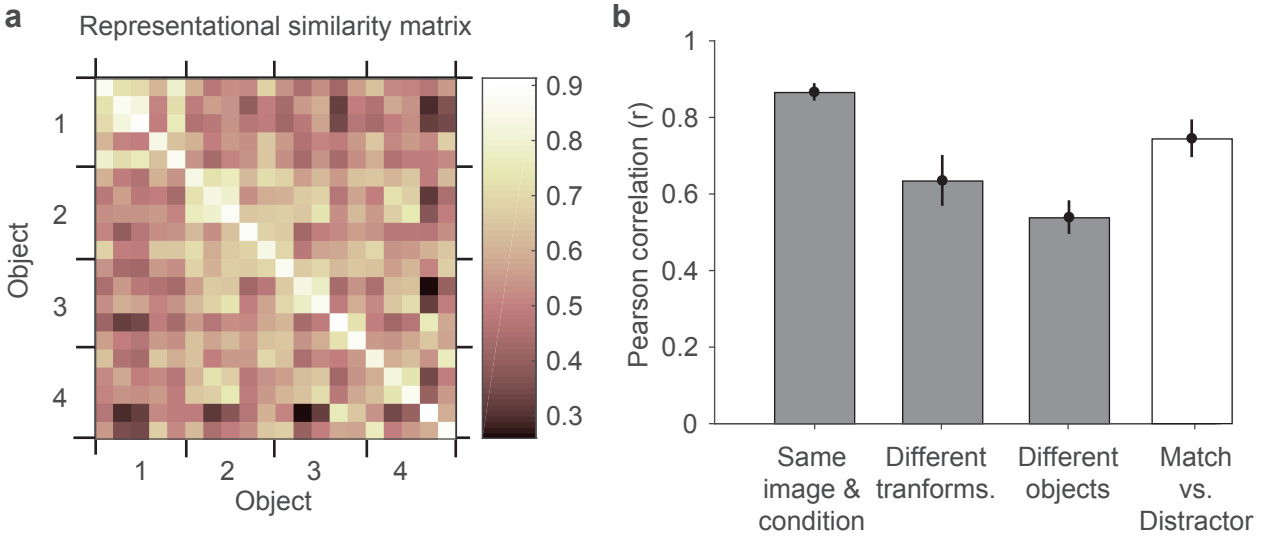
a
Representational similarity matrix
Object
1
2
3
4
Object
1
2
3
4
0.9
0.8
0.7
0.6
0.5
0.4
0.3
b
Pearson correlation (r)
1
0.8
0.6
0.4
0.2
0
Same image & condition
Different tranforms.
Different objects
Match vs. Distractor

Figure 7